



Selected papers from the
CLARIN Annual Conference 2022
Prague, Czechia



CLARIN

Selected Papers from the
CLARIN Annual Conference 2022

Prague, Czechia, 2022, 10-12 October

edited by Tomáš Erjavec and Maria Eskevich



Front Cover Illustration:

Picture Composition by CLARIN ERIC

Licensed under Creative Commons Attribution 4.0 International:

<https://creativecommons.org/licenses/by/4.0/>

Linköping Electronic Conference Proceedings

198

eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print)

2023

ISBN 978-91-8075-254-1 (PDF)

Introduction

Tomaž Erjavec

Programme Committee Chair
Jožef Stefan Institute, Ljubljana,
Slovenia

tomaz.erjavec@ijs.si

Darja Fišer

Executive Director of CLARIN ERIC
Institute of Contemporary History,
Slovenia

darja.fiser@clarin.eu

This volume presents the highlights of the eleventh CLARIN Annual Conference in 2022. The conference was held from 10 to 12 October 2022 as a hybrid event, with Prague, Czechia, as the venue.

CLARIN, the Common Language Resources and Technology Infrastructure, is a virtual platform that is accessible to everyone interested in language. CLARIN offers access to language resources, technology, and knowledge, and enables cross-country collaboration among academia, industry, policy-makers, cultural institutions, and the general public. Researchers, students, and citizens are offered access to digital language resources and technology services to deploy, connect, analyse and sustain such resources. In line with the Open Science agenda, CLARIN enables scholars from the Social Sciences and Humanities (SSH) and beyond to engage in and contribute to cutting-edge, data-driven research based on language data in a range of formats and modalities.

The infrastructure is run by CLARIN ERIC¹, a consortium of participating countries and institutes that was established in 2012 and has grown considerably in size since. Currently there are 22 member countries, 3 observers, and more than 100 associated research institutions, which are all encouraged and supported to be represented at the annual conference. The event is central for the CLARIN community and is one of the crucial instruments for CLARIN to function as a knowledge hub. At the conference, consortia from all participating countries and the various communities of use meet in order to exchange ideas, experiences and best practices in using the CLARIN infrastructure. Moreover, CLARIN2022 was also intended for the wider humanities and social sciences communities in order to exchange ideas and experiences within the CLARIN infrastructure. This includes the design, construction and operation of the CLARIN infrastructure, the data, tools and services that it contains or for which there is a need, its actual use by researchers, its relation to other infrastructures and projects, and the CLARIN Knowledge Infrastructure. Early in 2022 a call² was issued for which 21 abstracts were submitted. The authors of the submissions to the main conference sessions represented 11 CLARIN ERIC countries.

All submissions were reviewed anonymously by three reviewers (PC members and reviewers invited by PC members). Out of the 21 submitted abstracts, 16 submissions were accepted for presentation at the conference (acceptance rate 0.76). Those 16 submissions were grouped in the following subjects:

- Language Resources and CLARIN Centres (3 papers)
- Tools and Workflows (6 papers)
- Legal Questions (1 paper)
- Curation of Language Resources (3 papers)
- Research Cases (3 papers)

The accepted contributions were published in the online Proceedings of the Conference³.

As in 2018-2021, a PhD-session was organised as a combination of a one-minute pitch followed by a highly interactive poster session. The abstracts of 12 presentations were published in the online programme of CLARIN 2022⁴.

¹<http://www.clarin.eu>

²<https://www.clarin.eu/content/call-abstracts-clarin-annual-conference-2022>

³https://office.clarin.eu/v/CE-2022-2118-CLARIN2022_ConferenceProceedings.pdf

⁴<https://www.clarin.eu/content/programme-clarin-annual-conference-2022>

The 2022 edition of the CLARIN Annual Conference was shaped as a hybrid event. Live and on several screens, more than 300 registered participants were able to follow the quality content and learn what CLARIN is about. The conference programme contained both traditional conference elements, and novel items better suited to the virtual set-up:

- **Invited talk 'Enabling Digital Research - The German National Library as Part of a (National) Research Infrastructure.'** by Peter Leinen was followed by the **panel 'CLARIN and Libraries: Infrastructures Working Together'**, dedicated to further discussions on the topic which was chaired by Martin Wynne, with Sally Chambers, Andreas Witt and Peter Leinen as panelists. The invited talk presented the strategic orientation of the German National Library as well as the experiences and challenges faced by the library in concrete cooperation with the scientific community. Special attention was dedicated to the German National Library's commitment to the development of the National Research Data Infrastructure. In addition, the panelists discussed the outcomes of the 'CLARIN and Libraries' workshop that took place at KB National Library of the Netherlands in May 2022⁵, and focused on further steps that are to be taken in order to bring together the international CLARIN community and research libraries in order to deliver digital content for researchers.
- **Invited talk 'Is Human Label Variation Really so Bad for AI?'** by Barbara Plank. The talk elaborated on the question as to whether human variation in labelling is noise, or whether and how such information can be turned into signal for machine learning.
- **Invited talk '100 Years of Speech Recognition, the Data Fork and the Conversational Challenge. Stories from Today's Speech Industry'** by Ariane Nabeth-Halber. This talk showcased how several revolutions in the speech technology domain translate in the industry context, with a special focus on current advances that are starting to appear in the speech industry landscape, and outlined conversational challenge as the next frontier.
- **Panel 'CLARIN and Other SSH Platforms: You'll Never Innovate Alone'** was moderated by Jan Hajič with the following experts: Dieter Van Uytvanck, Alba Irollo, Simon Krek, Emilie Blotière and Sona Arasthe, and Matej Ďurčo. The panelists explained how they see their role in the SSH landscape and discussed the interoperability and complementarity of the different platforms.
- **Sessions of accepted conference papers** were organised as regular sessions with a presentation followed by Q&A.
- During the **CLARIN Student session**, PhD-students presented their work in progress: studies supported by or contributing to the CLARIN infrastructure. The aim of the session was to put the spotlights on the next generation of researchers and enable them to receive feedback on their work from CLARIN experts.
- The **Teaching with CLARIN session**⁶ invited university lecturers who had used CLARIN resources, tools or services in their courses to present their experience and suggest future steps that could help facilitate and accelerate the further integration of CLARIN into university curricula. Three of those submissions were granted with an award:
 - Ajda Pretnar Žagar, Kristina Pahor de Maiti, and Darja Fišer from the University of Ljubljana, Slovenia, for **'What's on the Agenda? Topic Modelling Parliamentary Debates before and during the COVID-19 Pandemic'**⁷
 - Jurgita Vaičenonienė from Vytautas Magnus University, Lithuania, for **'Lithuanian Collocations: Usage, Teaching, Learning, and Translation'**⁸

⁵<https://www.clarin.eu/event/2022/clarin-and-libraries>

⁶The slides of this and above mentioned CLARIN Students sessions can be found in the conference programme.

⁷<https://www.clarin.eu/content/whats-agenda-topic-modelling-parliamentary-debates-and-during-covid-19-pandemic>

⁸<https://www.clarin.eu/content/lithuanian-collocations-usage-teaching-learning-and-translation>

- Rachele Sprugnoli from Dipartimento di Discipline Umanistiche, Sociali e delle Imprese Culturali – Università degli Studi di Parma, Italy, for '**Natural Language Processing Methods**'⁹
- As usual, the **CLARIN Bazaar** provided an informal setting for conversations with CLARIN people and a space to showcase ongoing work and exchange ideas. The presenters were grouped together by topic to encourage interaction.
 - CLARIN Core
 - CLARIN National Nodes Highlights
 - Training and Knowledge Exchange
 - Data Curation Using NLP
 - EOSC-Related Activities
 - New Initiatives to Enhance Data and Services Coverage
 - New Collaborative Links
 - CLARIN Committees

After the event, CLARIN published a rich set of relevant materials:

- The complete conference programme and most of the slides presented: <https://www.clarin.eu/content/programme-clarin-annual-conference-2022>
- Recordings of keynotes, panels, and CLARIN Café that are available on the CLARIN YouTube channel: <https://youtube.com/playlist?list=PLIKmS5dTMgw0bGbaN3HSc9Dta4oBKqCxq>.
- Conference summary: <https://www.clarin.eu/content/clarin-annual-conference-2022-summary>

After the conference, authors of the accepted papers and student submissions, as well as participants of the Teaching with CLARIN session and CLARIN-funded projects, were invited to submit full versions of their papers to be considered for the post-conference proceedings volume. The papers were anonymously reviewed, each by three PC members. We received 13 (including 4 student papers) full-length submissions, and all of them were accepted for this volume. All the main topics addressed at the conference are covered in the papers.

We would like to thank all PC members and reviewers for their efforts in evaluating and re-evaluating the submissions, Maria Eskevich from the CLARIN Office for her indispensable support in the process of preparing these proceedings, and our colleagues at the Linköping University Electronic Press, who ensured that the digital publication of this volume came about smoothly. In order to support the programme chair and the programme committee in the organisation of reviewing and programme planning, a programme subcommittee was established, starting from CLARIN 2020. With respect to the establishment of the programme subcommittee, it was decided that the programme chair from the preceding year's conference should be one of the members in order to ensure continuity from one year's conference to the next. The members of the 2022 PC subcommittee were Tomaž Erjavec, Krister Lindén, Monica Monachini, and Koenraad De Smedt.

⁹<https://www.clarin.eu/content/natural-language-processing-methods>

Members of the Programme Committee for the CLARIN Annual Conference 2022:

- Starkaður Barkarson, Árni Magnússon Institute for Icelandic Studies, Iceland
- Lars Borin, University of Gothenburg, Sweden
- António Branco, Universidade de Lisboa, Portugal
- **Tomaž Erjavec, Jožef Stefan Institute, Slovenia (Chair)**
- Eva Hajičová, Charles University Prague, Czechia
- Marinos Ioannides, Cyprus University of Technology, Cyprus
- Arne Jönsson, Department of Computer and Information Science, Linöping University, SE-581 83, Linöping, Sweden
- Krister Lindén, University of Helsinki, Finland
- Monica Monachini, Institute of Computational Linguistics “A. Zampolli”, Italy
- Karlheinz Mörth, Austrian Academy of Sciences, Austria
- Costanza Navarretta, University of Copenhagen, Denmark
- Maciej Piasecki, Wrocław University of Science and Technology, Poland
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center, Greece
- Gijsbert Rutten, Leiden University, The Netherlands
- Kiril Simov, IICT, Bulgarian Academy of Sciences, Bulgaria
- Inguna Skadiņa, University of Latvia, Latvia
- Koenraad De Smedt, University of Bergen, Norway
- Marko Tadić, University of Zagreb, Croatia
- Jurgita Vaičenonienė, Vytautas Magnus University, Lithuania
- Vincent Vandeghinste, Instituut voor de Nederlandse Taal (Dutch Language Institute), the Netherlands & KU Leuven, Belgium
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary
- Andreas Witt, University of Mannheim, Germany
- Friedel Wolff, South African Centre for Digital Language Resources, North-West University, South Africa
- Joshua Wilbur, University of Tartu, Estonia
- Martin Wynne, University of Oxford, United Kingdom

Additional reviewer of this volume:

- Adriaan Lemmens, Belgium

Contents

Introduction <i>Tomaž Erjavec and Darja Fišer</i>	i
Analysing Changes in Official Use of the Design Concept Using SweCLARIN Resources <i>Lars Ahrenberg, Daniel Holmer, Stefan Holmlid and Arne Jönsson</i>	1
The CLaDA-BG Dictionary Creation System: Specifics and Perspectives <i>Zhivko Angelov, Kiril Simov, Petya Osenova and Zara Kancheva</i>	12
Linguistic Autobiographies. Towards the Creation of a Multilingual Resource Family <i>Silvia Calamai, Rosalba Nodari, Claudia Soria and Alessandro Carlucci</i>	23
The Pipeline for Publishing Resources in the Language Bank of Finland <i>Ute Dieckmann, Miitta Lennes, Jussi Piitulainen, Jyrki Niemi, Erik Axelson, Tommi Jauhiainen and Krister Lindén</i>	33
TEI and Git in ParlaMint: Collaborative Development of Language Resources <i>Tomaž Erjavec, Matyáš Kopp and Katja Meden</i>	44
EU Data Governance Act: Outlining a Potential Role for CLARIN <i>Paweł Kamocki, Krister Linden, Andrius Puksas and Aleksei Kelli</i>	57
Semantic Classification of Prepositions in BulTreeBank WordNet <i>Zara Kancheva</i>	66
Neural Metaphor Detection for Slovene <i>Matej Klemen and Marko Robnik-Šikonja</i>	77
Evaluation of the Archivio Vi.Vo Architecture: A Case Study on the Reuse of Legacy Data for Linguistic Purposes <i>Roberta Bianca Luzietti</i>	90
It-Sr-NER: CLARIN Compatible NER and Geoparsing Web Services for Italian and Serbian Parallel Text <i>Olja Perišić, Ranka Stanković, Milica Ikonić Nešić and Mihailo Škorić</i>	99
Lemmatizing and POS-tagging Akkadian with BabyLemmatizer and Dictionary-Based Post-Correction <i>Aleksi Sahala, Tero Alstola, Jonathan Valk and Krister Lindén</i>	111
Using classical readability formulas to measure text readability in Sesotho <i>Johannes Sibeko</i>	120
WebLicht-Batch – A Web-Based Interface for Batch Processing Large Input with the WebLicht Workflow Engine <i>Claus Zinn and Ben Campbell</i>	133

Analysing Changes in Official Use of the Design Concept Using SweCLARIN Resources

Lars Ahrenberg, Daniel Holmer, Stefan Holmlid, Arne Jönsson

Department of Computer and Information Science

Linköping University, Linköping, Sweden

firstname.lastname@liu.se

Abstract

We investigate changes in the use of four Swedish words from the fields of design and architecture. It has been suggested that their meanings have been blurred, especially in governmental reports and policy documents, so that distinctions between them that are important to stakeholders in the respective fields are lost. Specifically, we compare usage in two governmental public reports on design, one from 1999 and the other from 2015, and additionally in opinion responses to the 2015 report. Our approach is to contextualise occurrences of the words in different representations of the texts using word embeddings, topic modelling and sentiment analysis. Tools and language resources developed within the SweCLARIN infrastructure have been crucial for the implementation of the study.

1 Introduction

What is the relation between architecture and design? Should they be seen as concepts in the minds of speakers or as professions where stakeholders sometimes compete and sometimes join forces to achieve their goals? In this paper, we try to answer such questions using the resources developed for the analysis of Swedish by Språkbanken Text and distributed through the SweCLARIN portal. More specifically we want to study whether there is a change in the denotations and connotations of four related words: *arkitektur*, ‘architecture’, *form*, ‘form’, *formgivning* (cf. German *Formgebung*) and *design*. In particular, are there changes in their use and, perhaps, signs of a convergence? The study is limited to Sweden and Swedish as spoken in Sweden. Its rationale is a hypothesis from colleagues working in design that there has been an increased effort to place architecture and design under the same umbrella, not least from the side of the Swedish government, and that this development has been detrimental for the design field.

Dictionary definitions of the four words vary. According to one of them¹, the word *design* was first observed in Swedish in 1948. It is defined there as *konstnärlig formgivning*, ‘artistic form giving’ using the older term *formgivning*. It is not found in the Historical Swedish Dictionary, SAOB, as the words for the letter D were compiled and published at the beginning of the 20th century. Over the years *design* has established itself as a synonym of *formgivning*, and also, as will be shown, become the more frequently used of the two. The word *arkitektur*, ‘architecture’, on the other hand, is defined as a scientific discipline with a related concrete meaning as ‘artistic and technical design of buildings’. Thus, *arkitektur* can be defined in terms of *design* but also in other terms. Nowadays, also *design* can be studied at universities as a separate subject of study.

The word *form* has many meanings, one of them being ‘artistic form’. It is used in this sense by the private organisation *Svensk Form*, ‘Swedish Form’, established in 1845, and its journal, simply named *Form*. The mission of this organisation is to stimulate or inspire good design, and it aims to attract individuals and companies that work within the areas of “form, design and architecture”.

The main Swe-Clarín resources made use of in this study are the Sparv text analysis pipeline² (Borin

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

¹Nationalencyklopedins Ordbok, ‘The National Encyclopedia dictionary’, 1995 edition.

²<https://spraakbanken.gu.se/sparv/#/sparv-pipeline>

et al., 2016), the SenSALDO³ sentiment lexicon (Rouces et al., 2019), and the Swedish Culturomics Gigaword Corpus (Rødven Eide et al., 2016). In addition, we have used the Gensim framework (Řehůřek and Sojka, 2010) for word embeddings and topic modelling, and the VADER tools (Hutto and Gilbert, 2014) for sentiment analysis.

The paper is structured as follows. Section 2 gives a short background on the recent political developments and decisions relating to the fields of architecture and design. Section 3 provides an overview of our data and presents our general approach. In section 4 we present our results and, finally, in section 5 we state our conclusions.

2 Historical Background

In 1997 the Swedish government proposed an action program for an area identified as *arkitektur och formgivning*. Two years later an official governmental report (SOU), entitled *Mötesplats för form och design*, ‘A meeting-place for form and design’⁴ proposed a new initiative for design. Although the focus was on design the report argued that it was reasonable that there was a clear connection to architecture, as its proposals related to buildings and building sites and the main events were to happen in the upcoming ‘Year of Architecture’ referring to the year 2001.

A few years later, in 2009, the Swedish Museum of Architecture was given a new responsibility to cover also “other fields of design” and its name was changed to ArkDes: The Swedish Centre for Architecture and Design. Its mission is “to increase knowledge of and cultivate debate about how architecture and design affect our lives as citizens”⁵. The name ArkDes is derived from the two words *arkitektur* and *design* using the first three letters from each word. More recently a new SOU-report was requested which was ready in 2015. With the title *Gestaltad livsmiljö: en ny politik för arkitektur, form och design*, ‘Shaped habitat: a new policy for architecture, form and design’, it brought the three concepts architecture, form, and design closer together and proposed the establishment for a new public body dealing with them jointly. However, this has not yet been realised.

There are some qualitative studies of design policies, often comparing national contexts, but we are not aware of any study of a national design policy that makes use of similar computational text analytical models.

3 Method and Data

Ultimately, a close reading will be required to investigate how a certain concept such as architecture or design has been framed and understood in a set of documents. Distant reading of the kind we perform here can be useful, however, to catch general patterns in usage and provide quantitative estimates of them. We study the selected terms by term frequency counts and by comparing their local contexts. Contexts can be modelled in different ways each of which amplifies a different aspect or layer of meaning. We make use of three common models of context: word embeddings, topics, and sentiments. To generate these models we make use of open sourced software that is available on the GitHub platform. SweCLARIN resources are used for pre-processing, notably for parsing Swedish text and for supplying necessary lexical data on sentiments of Swedish words.

In addition to the two SOU reports mentioned above we have used the news sections of the Swedish Culturomics Gigaword Corpus (Rødven Eide et al., 2016) from relevant time periods for comparisons. Also included are 229 responses⁶ to the SOU report from 2015 to see to what extent they use the relevant terms in the same way as the report itself.

Responses expressing opinions on the proposals and general contents of SOU reports can be submitted by anyone. However, usually, a number of organisations with an interest in the subject matter of the report are invited to do so. The responding organisations can be categorised according to different criteria: being

³<https://spraakbanken.gu.se/resurser/sensaldo>

⁴SOU 1999:123

⁵<https://arkdes.se/en/about-us/>

⁶All SOUs and responses included in our analysis can be found at <https://github.com/holmad/Analysing-Changes-in-Official-Use-of-the-Design-Concept-Using-SweCLARIN-Resources>

from a Sector such as Private or Public (8 different), having a different Legal Status such as a Company, Interest organisation, Consultants, Municipality, Museum etc (17 different), or according to Area of Interest, such as Architecture, Research, Urban construction (18 different).

In the course of the study, we discovered that the terms of interest were often joined as conjuncts of a coordinate structure such as *arkitektur, form och design*, architecture, form, and design, or *arkitektur, formgivning och design*. In fact, in SOU 2015:88 it turned out that more than 50% of the instances of the words *arkitektur* and *design* occur as part of this coordination. For this reason, we decided to look at it as a concept of its own and created versions of the report where it is handled as a single token. In the sequel we will refer to it as the triad and the modified corpus will be referred to as the retokenized version, see Table 2.

Variant type	Examples
Base form	arkitektur, form och design
Uppercase first letter(s)	Arkitektur, form och design Arkitektur, Form och Design
Hyphenation	arkitektur, form och de-sign arkitektur-, form- och design
Misspellings	arkitektur, from och design
Misread PDF file	arkitektur, form och design,dnr
Definite variant + genitive	arkitekturen, formen och designen arkitekturens, formens och designens
Synonyms	arkitektur, formgivning och design arkitekturens, formens och gestaltningens
Compounds	arkitektur-, form- och designpolitiken arkitektur-, form- och designfrågor arkitektur-, form- och designområdena,

Table 1: Variants of the triad.

The triad actually appears in many different variants in the documents. Table 1 shows some of the most common variants. Some of these variants are mainly orthographic: the use of uppercase or lowercase, first letter of the first or all nouns, the inclusion of a hyphen in one of the words, or enclosing of it in quotation marks and other punctuation marks. Other variants are morphological using definite and/or genitive forms instead of the indefinite, nominative form. All of these could be caught by a regular expression. The triad may also be embedded in a compound where the second part is a noun such as *politik*, ‘politics’ and *område*, ‘area’. We decided to treat all orthographic and morphological variants as instances of the triad, excluding those where it is part of a compound. These are quite numerous, however, supporting the view that the triad has developed into a conceptual unit of its own.

Frequencies for the terms of interest in the different datasets are shown in Table 2. For the SOU reports and the responses to the 2015 report figures are given for both the original and the retokenized versions.

Corpus	arkitektur	design	formgivning	form	arkitektur, form och design	all tokens
Gw 1990-99 (News)	793	1,008	204	10,421	0	60,037,845
Gw 2010-2015 (News)	1,595	4,042	303	28,102	0	168,998,305
SOU 1999:123	43	358	139	180	0	42,263
- retokenized*	38	353	134	180	5	42,248
SOU 2015:88	318	328	26	301	0	34,739
- retokenized*	126	136	18	117	192	34,163
Responses to SOU 2015:88	2023	1659	29	1601	0	266,675
- retokenized*	1114	724	20	676	1311	262,742

*The triad *arkitektur, form och design* considered as one unit, see Section 4

Table 2: Frequencies of the investigated words in different corpora. The counts refer to tokens including orthographic and morphological variants.

As can be seen in Table 2 the two SOUs are roughly the same size, but the frequencies of the various words differ, as will be further discussed in Section 4. We also note that the total text of the responses to the 2015 SOU is much larger than the SOU itself.

4 Analyses

We have analysed the two SOUs and the 2015 Responses from four different perspectives. We first present results from a frequency count of the various terms, including the triad. We then present results considering the terms' use in general language and their semantic space by looking at word embeddings similarities. Finally we present results from topic analyses and sentiment analyses related to the terms.

4.1 Term Frequencies

The frequency of the terms is presented in Table 2, and illustrated in the bar charts of Figures 1 and 2.

A first observation is that the ratio of *design* to *formgiving* is changing rather rapidly. This is true for the news corpora where the ratio for the 1990s is about 5:1, rising to 13:1 for the period 2010-2015. The same holds for the SOUs where the ratio goes from about 2.5:1 in the 1999 SOU to close to 13:1 in 2015. Actually, the word *formgiving* is not only losing ground to *design* but also to *form*.

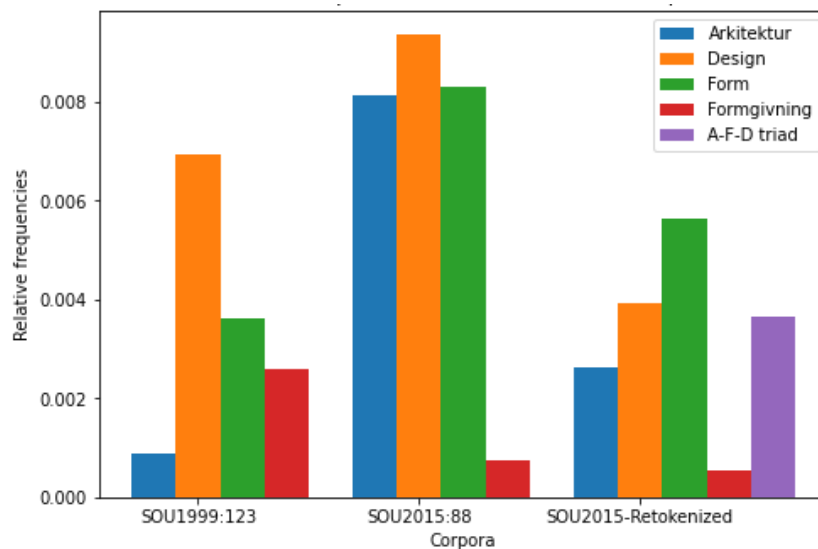


Figure 1: Relative term frequencies in the SOU reports.

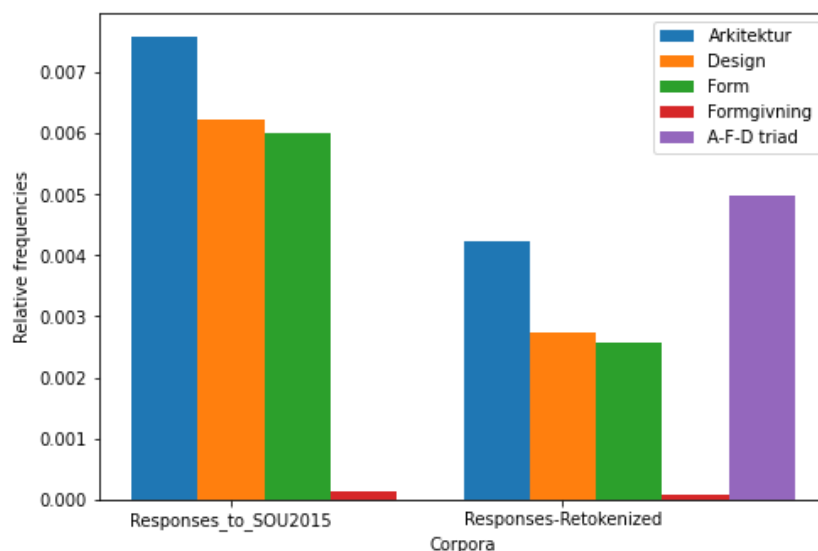


Figure 2: Effects of retokenization in the corpus of responses.

Secondly, we can see clearly the presence of the triad both in the report from 2015 and its responses. In fact, it is even more pronounced in the responses than in the report itself. See Figure 2.

4.2 Word Embeddings

To understand how the terms are used in general language, we looked at their distribution in the news sections of the Swedish Gigaword Corpus (Rødven Eide et al., 2016), for the periods 1990-99 and 2010-2015. Using several runs of word embeddings derived with the Gensim Word2Vec framework, we could observe the following:

- *design* and *formgivning* are close (synonyms) for both periods. The word *grafisk*, ‘graphical’, a common attribute to both terms, is about equally close.
- In the 1990s *formgivning* is a close neighbour to *arkitektur*, while *design* is further away. In the period 2010-15, the situation is reversed. In this period, *design* and *konst*, ‘art’ are competing for the place as the closest neighbour to *arkitektur*.
- The word *form* does not turn up in the close vicinity of any of the other words. This is due to its many other, more common meanings such as type, sort, shape, state, and mould.

Part of the reason why the words turn up as close neighbours in vector space is that they are often coordinated, in pairs such as *arkitektur och design*, but also in triples or even longer ones. Words that are frequent in these coordinations, apart from the four words under study, are *konst*, ‘art’, and *hantverk*, ‘crafts’. We also see trends of concept building via these coordinations, for example in the name of the Museum of Architecture and Design, ArkDes. These events in general language correspond well with the developments we see in the government documents and especially in the SOU from 2015, where the triad is so frequent.

We also produced word embeddings for the two reports. Due to the smaller size and the random character of word embeddings, these are harder to interpret. An interesting observation, however, is that in embeddings generated from the retokenized versions the neighbourhood of the triad does not include the individual terms, and vice versa.

In order to compare the semantic space of the studied terms in the two reports, we used the temporal word analogies method as suggested by (Szymanski, 2017). This method works by transforming two vector space models into a common vector space, which acts as a link between the models, enabling the comparison of word vectors between two otherwise independent models. Thus, we can investigate shifts between the models, in the form of “*which word X in model M1 correspond to word Y in model M2?*”.

We trained a Word2Vec model for each of the reports and applied the temporal word analogies technique to search for differences in usage of the studied terms, *architecture*, *design*, and *form*. Although the models themselves showed some differences when extracting and manually inspecting their most similar words, this method did not reveal any semantic shift of any of the studied terms between the two reports. It is possible, however, that this is due to the relatively small size of the data and vocabularies used.

4.3 Topic Modelling

We have applied topic modelling to the reports to see whether they differ in their distribution of topics, using the Gensim package (Řehůřek and Sojka, 2010) on parsed versions of the reports. Gensim provides an implementation of Latent Dirichlet Allocation (LDA) (Blei et al., 2003), widely used for topic modelling applications. The number of topics per model was chosen to maximise the coherence score C_v (Röder et al., 2015), which resulted in the model for the 1999 SOU having 16 topics, and the 2015 SOU having 14 topics.

We split each SOU according to its chapter. Since the chapters are few but lengthy, we performed a second splitting process where each chapter was divided into chunks of ≈ 100 tokens. Each of these chunks was seen as a document in the topic modelling process. We tokenized, lemmatized and POS-tagged each document with the Sparv pipeline. This allowed us to only include words in their lemmatized form, as is fairly common practice, and to only include content words (nouns, adjectives, verbs, and adverbs), which should make the topics more interpretable.

We could see for the 2015 SOU, that for the topics where *design* is among the 10 most relevant terms, so is *arkitektur*, and vice versa. For the majority of topics where this happens, *form* is also among the 10 most relevant terms.

In addition to the topic modelling of the SOUs, we trained topic models also on the responses to the 2015 SOU. This was done to enable studies of the topic distributions on different categories of responding organisations. For this task, we used the BERTopic library (Grootendorst, 2022) that leverages the recent years’ rise of pre-trained transformer-based language models and is able to produce topic models based on the semantic structure (rather than only the word frequencies) of a collection of documents. The library utilizes the pre-trained language model (in our case, we used a Swedish sentence-transformer model (Rekathati, 2021)) to create document embeddings, cluster the embeddings, and through a class-based TF-IDF procedure generate topics.

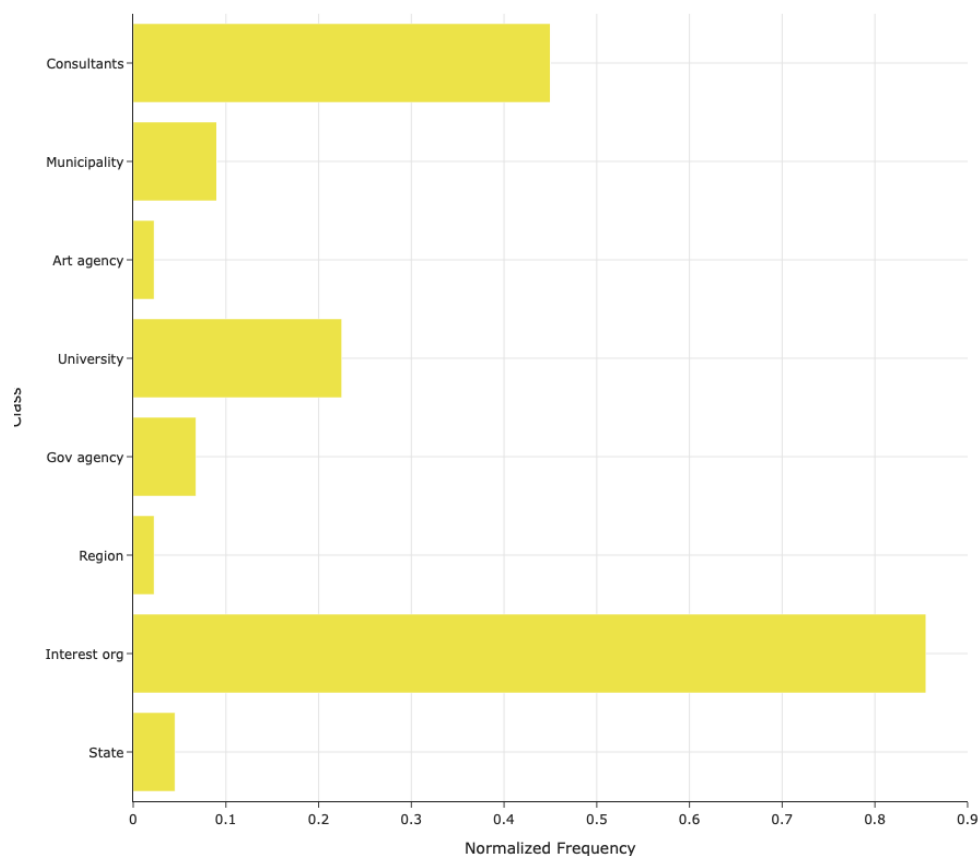


Figure 3: Relative frequency of a topic related to the term *design* across responses from different category classes.

The topic model was trained from all the SOU 2015 responses in our dataset. Within the BERTopic-framework we used the default UMAP model for dimensionality reduction and HDBSCAN for clustering (both with default parameters). The resulting model consists of 106 topics. Since all responses were assigned a category of origin, we could utilize the support BERTopic provide to visualize topics per predefined class.

In Figure 3, the topic most heavily related to the term *design*, including the word *design* itself and compounds such as *designmetodik*, ‘design methodology’, *designkompetens*, ‘design competence’, etc, are displayed to have the most prevalence in responses from respondents assigned to the legal status Interest organizations and Consultants. This design topic was highlighted only by organisations from 8 of the 17 legal statuses. From the wider range of actors, this topic did not emerge at all.

A topic heavily related to the term *arkitektur*, including words like *bostäder*, ‘residences’, *människor*, ‘humans’, and *förtätning*, ‘urban consolidation’, can be observed to just like *design* have a high presence in responses from Interest organizations as shown in Figure 4. This topic, however, is also often located in responses from Municipality, University, and Government Agency actors. Many categories of actors

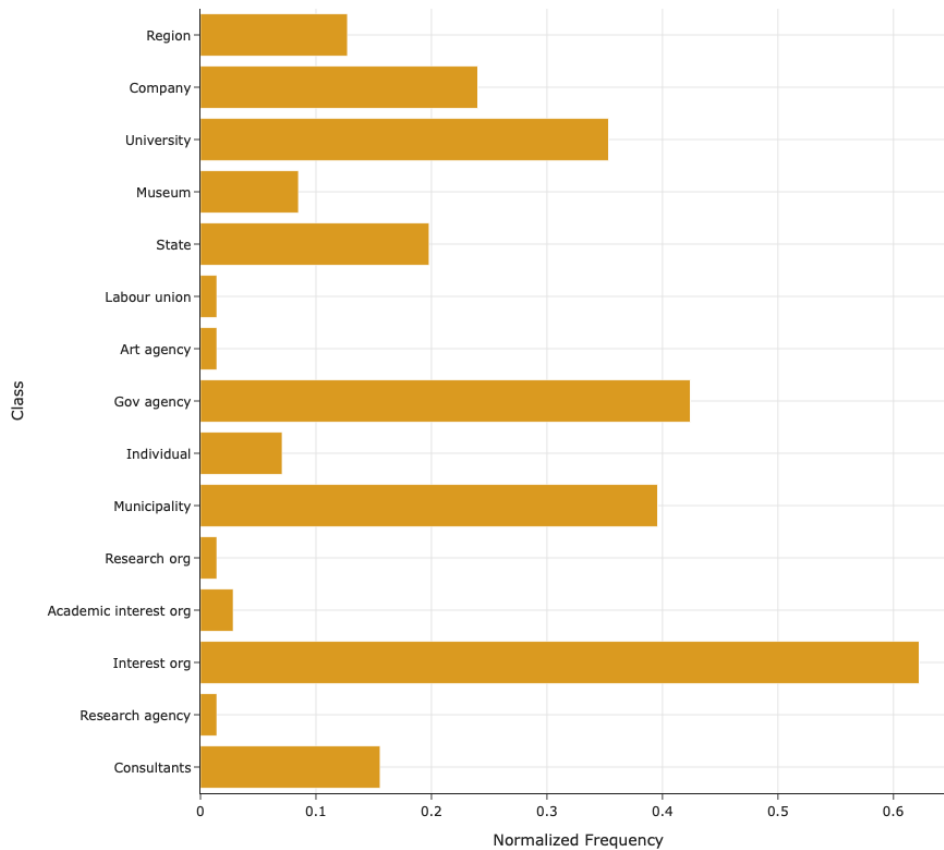


Figure 4: Relative frequency of an *arkitektur*, 'architecture', related topic across responses from different category classes.

find it important to mention this topic in their responses, and trivial visual inspection shows that the ones highlighting it the most are the urban planning and commissioning actors as well as part of the businesses engaged. However, it is not possible from this analysis to see whether they agree with each other, whether they agree with the SOU or whether they disagree with the SOU.

4.4 Sentiment Analyses

Sentiment analysis was applied to give another perspective on the reports. In particular, we were interested to see whether there is a change in the way the relevant terms are presented in the two reports.

For sentiment analysis we used Vader⁷ (Hutto and Gilbert, 2014) and the Swedish SenSALDO 0.2 sentiment lexicon (Rouces et al., 2019). The lexicon in English Vader comprises 5500 lexical entries with sentiment scores between +5 and -5. There is a Swedish version with a lexicon comprising 2067 lexical entries and sentiment scores +3 and -3, but less granular (Borg and Boldt, 2020). SensSaldo uses the three sentiment scores -1, 0 and +1. What makes SenSALDO unique in a Swedish context is that it assigns different sentiment values to different senses of a word, for instance, the Swedish word *fara* can mean 'danger' or 'go (away)' where the former has a negative sentiment and the latter is neutral. SenSALDO comprises 12287 lexical entries where 8893 are unique words. Word sense disambiguation with the SenSALDO 0.2 lexicon is made possible by using the Sparv pipeline. Vader also uses booster words, such as *amazingly*, to further refine the sentiment analysis. The booster dictionary used in the analyses is a slightly enhanced version of the dictionary used for sentiment analysis of e-mail conversations (Borg

⁷<https://github.com/cjhutto/vaderSentiment>

and Boldt, 2020) and comprises 89 items.

Vader produces a compound score for each sentence, by summing the valence scores of the words according to their identified sense and normalising this sum to be between -1 and +1. It is also useful to calculate the amount of positive, negative or neutral sentences. For this, we use the recommendations that a sentence has positive sentiment if the compound score is ≥ 0.05 , neutral if the compound score is > -0.05 and < 0.05 and negative if it is ≤ -0.05 ⁷.

Generally speaking, we find the sentiment score produced by Vader intuitively correct, although it has some problems with negations. A typical sentence with a positive sentiment from the responses, with a compound score of 0.8402 is: “*Med bättre kunskap och medvetenhet om vad god arkformdes innebär för människors välbefinnande kan vi skapa processer som långsiktigt främjar en god samhällsekonomi och goda livsmiljöer för alla*” ‘With better knowledge and awareness of what good arkformdes means for people’s well-being, we can create processes that promote a good social economy and good living environments for everyone in the long term’.

A negative sentence with a compound score of -0.6124 is “*Kritiken har i hög grad skjutit in sig på brister i ledningen av verksamheten, men i själva verket har nog uppgiften att vidga tidigare Arkitektmuseets uppgifter till att även omfatta design varit olämplig*”, ‘The criticism has largely focused on shortcomings in the management of the business, but in fact the task of expanding the previous tasks of the Museum of Architecture to also include design has probably been inappropriate’.

And, finally, with sentiment score 0, a typical neutral sentence is “*Ett sådant exempel är de strukturer som förekommer i södra Sverige, med samarbete mellan bland andra Region Skåne, Malmö Stad, Form/Design Center, regionala branschplattformar, högskolor och universitet*”, ‘One such example is the structures that exist in southern Sweden, with collaboration between, among others, Region Skåne, Malmö City, Form/Design Center, regional industry platforms, colleges and universities’.

SOU	Sentences	Negative Sentences	Positive Sentences	Neutral Sentences	Mean sentiment
1999	2161	118 (6%)	820 (38%)	1223 (56%)	0.112*
2015	1581	87 (6%)	722 (49%)	722 (45%)	0.168*

Table 3: Descriptive statistics. *Significant, $p < 0.01$

SOU	Form		Formgivning		Arkitektur		Design	
	Sentences	Sentiment	Sentences	Sentiment	Sentences	Sentiment	Sentences	Sentiment
1999	180	0.160*	139	0.125	43	0.176*	358	0.141*
2015	301	0.275*	26	0.254	318	0.288*	328	0.248*

Table 4: Concept-based sentiment for the SOUs from 1999 and 2015 in original versions. The number of sentences and mean concept sentence sentiment. *Significant, $p < 0.01$

The results from analysing the two SOUs are depicted in Table 3. In both SOUs the amount of sentences having a negative sentiment is small, roughly 6% each year. The main difference is that the number of positive sentences is higher in 2015 than in 1999, and, correspondingly, the amount of neutral sentences is lower in 2015 compared to 1999. This can be seen in the mean sentiment. For the 1999 SOU, it is 0.112 and for 2015 it is 0.168. The difference is significant⁸, $p < 0.01$ ⁹. Thus, the 2015 SOU uses overall a more positive tone.

We have also investigated the sentiment for each of the concepts in focus *arkitektur*, *design*, *form*, and *formgivning*. For each concept, sentiment is computed if the concept occurs in the sentence. The results are presented in Table 4 showing that the text in the 2015 SOU has a more positive attitude towards the concepts *form*, *arkitektur* and *design*. The difference in sentiment for *form* between 1999 and 2015 is significant¹⁰. The difference in sentiment for *design* between 1999 and 2015 is significant¹¹, and the

⁸Welch’s t-test = 7.2511, $p = 0.0000$.

⁹For all further significance tests we use $p < 0.01$ to denote a significant difference.

¹⁰Welch’s t-test= 4.8204, $p = 0.000$

¹¹Welch’s t-test = 5.7667, $p = 0.000$

difference for *arkitektur* is significant¹². The difference for *formgivning* is not significant for $p < 0.01$, but $p = 0.0230$ so there is a tendency.

SOU	Form		Formgivning		Arkitektur		Design		triad	
	Sents	Sentiment	Sents	Sentiment	Sents	Sentiment	Sents	Sentiment	Sents	Sentiment
1999	180	0.160	134	0.125	38	0.174	353	0.140	5	0.192
2015	117	0.239	18	0.172	126	0.266	136	0.170	192	0.304

Table 5: Concept-based sentiment for the SOUs from 1999 and 2015 in retokenized versions. The triad is a separate concept. The number of sentences and mean concept sentence sentiment.

If we consider the triad *arkitektur, form- och design* and use the retokenized corpus to filter out all sentences containing it, see Table 5 for descriptive statistics, there are no significant differences.

The triad is not used much in 1999, only 5 occurrences, so we also compared the triad to the other concepts for 2015, last row in Table 5, and then it turns out that only the difference in sentiment for *design*, 0.168, compared to the triad, 0.304, is significant, $p < 0.01$. That is, the triad is presented with a more positive sentiment than *design* in the 2015 SOU.

Corpus	Form		Formgivning		Arkitektur		Design	
	Sentences	Sentiment	Sentences	Sentiment	Sentences	Sentiment	Sentences	Sentiment
Retokenized	117	0.239	18	0.172	126	0.266	136	0.170*
Not retokenized	301	0.275	26	0.254	318	0.289	328	0.248*

Table 6: Concept-based sentiment for the SOU 2015 filtered for the triad or not. The number of sentences and mean concept sentence sentiment. *Significant, $p < 0.01$

In Table 6 we compare the last rows of Table 4 and Table 5 without the triad column to see whether there are sentiment differences for the four concepts considered. Again the only significant difference is for *design*¹³.

Corpus	Form		Formgivning		Arkitektur		Design		triad	
	Sents	Sentiment	Sents	Sentiment	Sents	Sentiment	Sents	Sentiment	Sents	Sentiment
Responses	676	0.171	20	0.071	1114	0.180*	724	0.142	1311	0.224*
SOU 2015	117	0.239	18	0.172	126	0.266*	136	0.170	192	0.304*

Table 7: Concept-based sentiment for the 2015 SOU and the responses with the triad as a separate concept. The number of sentences and mean concept sentence sentiment. *Significant, $p < 0.01$

The sentiment analyses of the responses to the 2015 SOU are presented in Table 7. The difference for the concept *arkitektur* in the SOU, 0.266, and the responses, 0.180, is significant¹⁴, and the difference for the triad SOU = 0.304, responses = 0.224, is significant¹⁵, i.e. the concepts are used more positively in the SOUs compared to the responses.

5 Conclusions and Reflections

By using a variety of methods and SweCLARIN resources we were able to present textual analyses of the material from different perspectives. The tools and language resources available in the SweCLARIN infrastructure for analysis of Swedish texts enable comparisons of language use also over such short time spans as 20 years. In particular, we exploited the ability of the Swedish SenSALDO lexicon to identify word senses for sentiment analysis, and the Culturomics Gigaword Corpus for comparing official government reports with general language. Most of our analyses, the sentiment analysis, the topic modelling and the frequency calculations, utilized texts parsed using the SweCLARIN Sparv pipeline.

Although the SweCLARIN resources are in most cases straightforward to use, some fine-tuning of Sparv was needed. It would not have been possible to do the analyses without further programming, to

¹²Welch's t-test = 3.5494, $p = 0.0007$

¹³Welsh's t-test = 3.2559, $p = 0.0013$

¹⁴Welsh's t-test = 3.5261, $p = 0.0006$

¹⁵Welsch t-test = 3.7176, $p = 0.0002$

prepare texts for analysis using regular expressions, and to adopt libraries, such as Vader, to the language resources. For modelling of context we used open source software available on other platforms for sharing resources such as GitHub. Simplifying somewhat we can say that SweCLARIN supplied the necessary language-specific tools and lexical resources while the language-independent modelling software was obtained elsewhere. We think that researchers who want to apply language technology tools to their problems need to be aware of several repositories and what they can offer. Guides for their use, and particularly, when you need to combine them with language-specific data and pre-processing, can be part of pedagogical material and the handbooks provided by CLARIN centers. SweCLARIN, as an example, maintains such a handbook¹⁶.

We have analysed and compared two public government reports, both being part of a political process of policy development, suggesting how financial and structural support should be developed for the areas architecture, design and form giving. The first was published in 1999 and the second in 2015. We have also included responses to the 2015 report in the analyses. Sentiment analysis seems to be the method that provides the most reliable results in our case, while the results from topic modelling and temporal word analogies are more uncertain due to the small dataset. However, with the BERTopic library, we could relate topics to metadata of the responses.

On the one hand, we can see the word *design* being used with increased frequency overtaking the role of the older word *formgivning*. This reflects a general shift in language usage. On the other hand, when we look at the relation between the words *design* and *arkitektur*, we can see indications of semantic convergence, especially in the latter report. This is not least due to the creation of a vague super concept expressed by the coordinate structure *arkitektur, form and design*, which we refer to as the triad. In the former report, the triad is very uncommon. In the latter report either the triad or one of the individual terms is used, indicating that the suggestion of the report is to establish this super concept as a single area of policy. The triad, which is so frequent in the 2015 report, is significantly more positively described there, compared to *form* and *design* as individual terms, seemingly underlining a supporting policy that should take an integrative approach. Moreover, the triad is used in more positive terms than *design*, while the sentiment of the term *arkitektur* is not as influenced by the usage of the super concept. This indicates a possibility that there is bias in the report, or that there are more issues to deal with in the area of design than in architecture.

In the responses to the report from 2015, the triad is reused, and taken onboard as a concept to comment on. Sentiment analysis shows that the super concept is used more positively in the report than in the responses, indicating that there may be criticism in the responses toward the conflation of the three areas. Moreover, in the responses, few topics were related to design, and only a small number of actors were responding about design. This indicates that the majority of respondents were not reacting much to how *design* was handled in the report.

Overall, the analyses make it possible to identify potential aspects that could be further analysed for policy development. Structural issues, such as how certain areas are handled in relation to others, or what topics different clusters of actors are engaged in, can be identified. Also, content issues can be identified, such as how much focus different topics are given, but also more specifically how different interest areas are populated.

References

- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.
- Anton Borg and Martin Boldt. 2020. Using VADER sentiment and SVM for predicting customer response sentiment. *Expert Systems with Applications*, 162:113746.
- Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, and Anne Schumacher. 2016. Sparv: Språkbanken’s corpus annotation pipeline infrastructure. In *SLTC 2016. The Sixth Swedish Language Technology Conference, Umeå University, 17-18 November, 2016*.

¹⁶<https://sweclarin.se/swe/handbok>

- Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Ann Arbor, MI.
- Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. <http://is.muni.cz/publication/884893/en>.
- Faton Rekathati. 2021. The KBLab blog: Introducing a Swedish sentence transformer. <https://kblabb.github.io/posts/2021-08-23-a-swedish-sentence-transformer/>.
- Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM '15*, page 399–408, New York, NY, USA. Association for Computing Machinery.
- Jacobo Rouces, Nina Tahmasebi, Lars Borin, and Stian Rødven Eide. 2019. Sensaldo: Creating a sentiment lexicon for Swedish. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pages 4192–4198.
- Stian Rødven Eide, Nina Tahmasebi, and Lars Borin. 2016. The Swedish Culturomics Gigaword Corpus: A one billion word Swedish reference dataset for NLP. In *Digital Humanities 2016. From Digitization to Knowledge 2016: Resources and Methods for Semantic Processing of Digital Works/Texts, Proceedings of the Workshop, Krakow, Poland*.
- Terrence Szymanski. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 448–453, Vancouver, Canada, July. Association for Computational Linguistics.

The CLaDA-BG Dictionary Creation System: Specifics and Perspectives

Zhivko Angelov, Kiril Simov, Petya Osenova, Zara Kancheva

Institute of Information and Communication Technologies

Bulgarian Academy of Sciences

Sofia, Bulgaria

angelov.zhivko@gmail.com, {kivs,petya,zara}@bultreebank.org

Abstract

The paper reports on the current status of a system for creating dictionaries within the CLaDA-BG infrastructure. The system is called CLaDA-BG-Dict. At the heart of the system lies the lexical thesaurus BTB-Wordnet around which all other language resources for Bulgarian are organized. These are various types of dictionaries (morphological, explanatory, terminological, etc.), ontologies (such as DBpedia), corpora (in-house and external). The specific features and functionalities of the system are discussed with respect to the language resource integrity. Also, the rationale behind the construction of such a system are given together with an outline of its utility for a number of NLP tasks and for various types of users. The ideas presented as well as the system itself are scalable to integrating resources also for other languages.

1 Introduction

In this paper we present the main principles and perspectives behind the CLaDA-BG Dictionary Creation System — **CLaDA-BG-Dict**. The ultimate goal of its implementation is to support the compilation of new dictionaries by individuals or collaborators with respect to a certain task and through the usage of all the available resources within the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH — CLaDA-BG¹.

We aim to provide a system that supports the whole cycle of creating various types of dictionaries. At the heart of this system lies the Bulgarian BulTreeBank WordNet (BTB-WN) — (Osenova and Simov, 2018). It has been developed as an aggregator of semantic knowledge around which other dictionaries and sources of information (including grammatical, encyclopedic, etc.) have been organized in the form of a(n) (inter)linked knowledge network.

The motivation for the development of the CLaDA-BG-Dict refers to the need for: better control on the consistency in the creation of lexical language resources; user friendly and communicative collaborative environment; better connections among the available resources. Also in the light of open data we expect that there will be more lexicographical data available for reuse in future. This would facilitate the rapid creation of specialised lexicons as well as their publishing and focused usage.

The incentive for the design and implementation of CLaDA-BG-Dict system was the development of BTB-WN. On the one hand, we were aware that there already exist software systems for the creation of other wordnets such as BulNet, GermaNet and Polish Wordnet (plWordNet). However, these systems inevitably reflect the approaches of the creators of these wordnets and thus, they do not support all the functions, needed for the work on BTB-WN. These are: extension of lexical entries structure; mapping to other resources (inflectional lexicons, explanatory dictionaries, bilingual dictionaries, Wikipedia pages, etc.); concordance for selection of examples; ticketing system for identifying and handling errors of various types and history of changes. On the other hand, the workflow on a contemporary wordnet requires

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

¹<https://clada-bg.eu/en/>

the addition of linguistic information beyond synsets, lemmas, definitions and relations. Such information includes: links to grammatical paradigms, word valencies, links to Wikipedia, mappings to other wordnets, appropriate examples, etc.

For all the reasons presented above, in this paper we present our customized solution for a resource integrating tool. The idea and implementation are scalable also to resources in other languages.

The structure of the paper is as follows: the next section provides a focused review of related works. Section 3 discusses the specifics and functionalities of **CLaDA-BG-Dict** system. Section 4 outlines the language resources that support dictionary creation. Section 5 concludes the paper and presents plans for future work.

2 Related Work

In our work we follow the approaches described in two existing wordnet editing systems —(Henrich and Hinrichs, 2010) and (Naskret et al., 2018). Similarly to Henrich and Hinrichs (2010) we needed to switch from a tool that supported only local editing where synsets were considered within a very limited context to a tool that supports editing of the wordnet data within a larger context. Comparably to both systems we switched from a file oriented presentation of data to a centralized database used via web to support simultaneous work of a team of experts. The most important benefit of this switch is that each member of the working team started to observe the changes made in content and structure at the time they had been made. In addition, the user of the system has the possibility to consult the resources in the database in any moment when this is necessary. Thus, the users have at their disposal a global view over the wordnet. As a consequence, when editing, they might take into account all the data in a connected way instead of partial or isolated views.

Here the following question might arise: Why to develop yet another system when there are already so many out there? We decided to implement our own system because in addition to developing our wordnet, we wanted to support and connect all the language resources we already had incorporated within it. These are: a spelling and grammar dictionary, an explanatory dictionary, bilingual dictionaries, related corpora for providing adequate examples that register various characteristics of the respective meaning. For us the mapping between the existing language resources is set as an important goal. Thus, we wanted to support such mappings as early in the process of the wordnet creation as possible.

Our aim is to extend the current system further towards a full-fledged dictionary writing system. It is envisaged to provide the necessary environment and services for the compilation of ad hoc and task-oriented dictionaries through the access to all the language data – starting from the existing dictionaries, corpora, encyclopedic knowledge, and others.

We are aware that many efforts have already been invested in dictionary creation systems from various points of view: formats and standards; approaches in the representation of the linguistic knowledge; implementation strategies, etc. Here we mention only some of the related work. One of the most influential ongoing frameworks is ELEXIS². After having performed an in-depth survey on the needs of lexicographers³ — (Kallas et al., 2019), the team behind ELEXIS (p. 62) envisaged ‘two complementary sets of tools will be provided: lexicographic workflow tools and crowdsourcing and gamification tools. The first will include a user friendly open-source online dictionary writing system, with the aim to provide the central dictionary writing platform for new lexicography which also includes new possibilities of online collaboration. The other will provide tools for new techniques of dictionary creation, such as explicit or implicit crowdsourcing (gamification).’ There are two tools for dictionary creation provided by ELEXIS – OneClick Dictionary⁴ and Lexonomy⁵ — (Měchura, 2017). The OneClick Dictionary is a dictionary drafting module, a feature of Sketch Engine⁶ which produces a machine generated dictionary draft that is later edited by lexicographers in the Lexonomy module. Functionalities such as wordlists,

²<https://elex.is/>

³https://elex.is/wp-content/uploads/2019/02/ELEXIS_D1_1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf

⁴<https://github.com/elexis-eu/ocd>

⁵<https://lexonomy.elex.is/>

⁶<https://www.sketchengine.eu/>

corpora, concordance, thesauri, etc. are also integrated in the tools. Similarly to the CLaDA-BG-Dict system, these tools are applicable also for the tasks of creating glossaries and domain-specific wordlists and dictionaries. Unfortunately, when we started the implementation of CLaDA-BG-Dict system, these tools were not available for a public use. Thus, we plan to customize and adapt them to our framework as much as possible in our future work.

Our current system supports Lexical Markup Framework (LMF) formats but not in its full capacity. LMF files can be uploaded, edited and then saved outside the system. However, not everything from LMF is supported. There is no converter from the internal files into Lemon Standard and back⁷. At this point we rely only on the LMF-based converters. It should be noted once again that we aim to facilitate the work not only of the professional lexicographers but also of any other researcher groups and common users. Thus, we imagine helping teachers to compile a dictionary of minimum words/senses, etc. for their class; or a student to construct incrementally a learner lexicon of Bulgarian related to a language that they know, etc. Within DARIAH-ERIC a standard for representation of dictionaries has been developed — TEI-Lex0.⁸ This standard is supported also by ELEXIS. Since it has already been established as a best practice, we plan to use it as well. At the moment we support only the minimum to exchange data in BTB-WN whereas the complete set of import and export formats needs to be implemented. The main focus in implementing the system was put initially on the availability and integration of the resources. Thus, our efforts on ensuring adequate exchanging formats and adherence to the common standards come next.

Last but not least, one of the CLARIN-ERIC Resource Families are Lexica. They are 89 and most of them are monolingual.⁹ They are of various types – inflectional, morphological, valency, multiword, stopwords, sentiment, etc. Thus, they are a good source for insights in adding more types of resources and more types of analyses into the system.

3 System Specifics and Functionalities

Initially, CLaDA-BG-Dict was designed and implemented to support the verification and extension of BTB-WN. The motivation for this was that the existing version of BTB-WN was initiated in an XML format within the CLaRK System¹⁰ — (Simov et al., 2001). The XML format used during the creation of earlier versions of BTB-WN was not a standard one. It was designed to facilitate the editing of lexical entries for each synset. Also this XML format had to reflect the incorporation of non-standardised data such as links to Open English Wordnet (OEW), Bulgarian Wikipedia, and others. However, the creation of BTB-WN in this way revealed some shortcomings. As mentioned above, the main problem with working in CLaRK System was that the users had only a local view over the existing Bulgarian synsets because the data with BTB-WN were stored in several XML files, and searches had to be performed within each of them (or in some of them). For instance, it was not easy to observe all the synsets in which a given lemma participates, because they could be in different XML files. Thus, one of the main design solutions was to support the mapping to the OEW with the idea to enhance the multilingual applications and the transfer of information from OEW to BTB-WN. In addition, we needed some system support for the better integration of BTB-WN with other language and knowledge resources for Bulgarian.

The system is a client-server web-based editor using a thick client model. The thick client is installed on the user computer (desktop or laptop). The thick client as a user interface to the system provides a better flexibility with respect to implementing the necessary functionality. It especially facilitates the way to compose several actions during the creation and editing of new synsets, assigning shortcuts and others.

3.1 Initial Acquaintance with the System

The database is installed on a server and it is accessed online via the web. A relational database is used for storing the data. Two people are not allowed to work on the same synset at the same time. They can

⁷<https://www.w3.org/2019/09/lexicog/>

⁸<https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>

⁹<https://www.clarin.eu/resource-families/lexical-resources-lexica>

¹⁰<http://bultreebank.org/en/clark/>

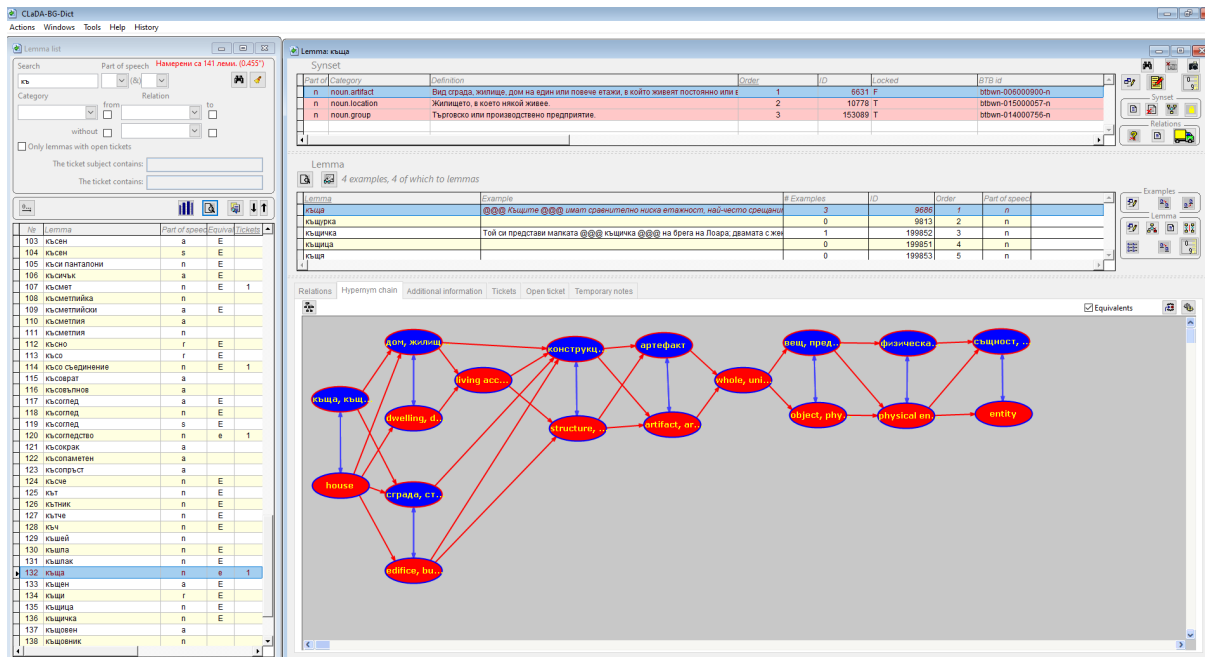


Figure 1: A screenshot of the user interface of CLaDA-BG-Dict.

work subsequently on the same synset or simultaneously on different synsets. Also, the system stores in a log file the following information: each editing step, the name of the person who edited, and at what time the edit was done. In this way we might track back states of data and repair errors if necessary. The system reflects our approach towards the next wordnet developments as well as towards the integration of various language resources within any dictionary compilations. This approach reflects the lexicon-grammar interface in a better way. In Fig. 1 a screenshot is presented, which shows a search over lemmas starting with “ка” — this search string is not a word in Bulgarian and only serves as a query which selects a range of lemmas within BTB-WN starting with it. The search string and the result from the search are displayed within the left element of the window in the figure, named *Lemma list*. This part of the window is separated in three parts. The upper part supports searches in the database. Searches can be performed by several criteria: by string, by POS or category, by relations or by the type of the ticket that was assigned to a synset. The list of results from the search is presented in the bottom part of the left element of the window. Each row in this table contains a lemma and the POS of the lemma. It gives information whether there is an equivalent synset in OEW, the number of tickets it has, etc. The search with the string “ка” returns more than 140 lemmas, among which “къща” *kashta* (‘house’). The middle part contains icons for the possible operations over lemmas in the list like - sorting, statistics and opening of a selected lemma within an editor form.

Thus, when the lemma *къща* is selected (as in the figure) and the editor form is opened (displayed on the right side of the window), it can be seen that there are three meanings (synsets) with the categories *noun.artifact*, *noun.location*, and *noun.group* which contain this lemma. The editor form is related to a given lemma (marked in the left upper corner — Lemma: *къща*). In this way the system allows for the simultaneous opening of several editor forms. These might be used to support the comparison of different synsets for different lemmas. Also they might facilitate the creation of relations between various synsets.

Each editor form consists of three areas: *Synset* area, located at the top part of the form, *Lemma* area, located in the middle part of the form, and *Miscellaneous* area at the bottom of the form. The *Synset area* contains information for the synsets that include the lemma related to the editor form. The information of a given synset includes: the category of the synset; the definition; the order of synsets for the lemma, the

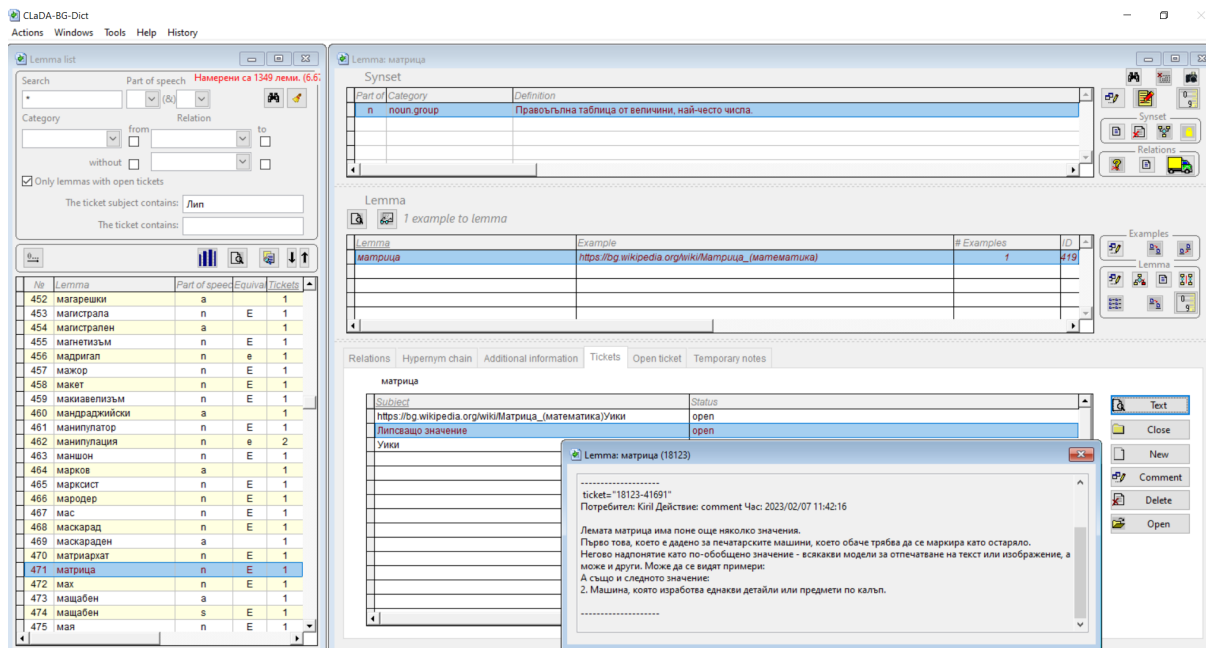


Figure 2: A screenshot of the tickets in CLaDA-BG-Dict.

internal ID of the synset (unique and unchangeable within the database), the locking information, and the BTB-WN ID of the synset. Next to the table are the operations that can be performed over a given synset. These include: editing of the definition (if not locked), export of the synset in a textual form, reordering of the synsets of a lemma (see below), creation of a new synset, deletion of a synset, manipulation of relations. The *Lemma area* consists also of two elements – a table and a tool pane. In the table a list of lemmas in the synset is presented. Each lemma is associated with some examples for the given synset, its paradigm and internal ID. The lemmas in the same synset are ordered with respect to how well they represent the meaning of the synset. The lemma tools include manipulation of examples, editing of lemmas, editing of the paradigms, and ordering of the lemmas. The *Miscellaneous area* consists of several tabs presenting different parts of information like relations for the synset, hierarchy of synsets, information from associated machine readable dictionaries, tickets, and temporal notes on some information in the synsets. In the above figure – in the Miscellaneous area – a graphic representation of a hierarchy of a noun is given and also the mapping to the English synsets. The graphic of relations shows the hierarchy of the first synset – noun. artifact. It is a hyponym of *building*, *construction*, *artefact*, and on the highest level – of *physical entity* and *entity*.

Regarding the BTB-WN, the system provides information about a selected lemma: its meanings (synsets) and associated examples; its internal relations as well as the mappings to the OEW; it also provides the ratio among the used relations. In case of equivalent synsets between BTB-WN and OEW, the Bulgarian synset inherits all the relations from the English synsets. In cases when these equivalent synsets have also corresponding hypernyms or hyponyms, the inherited relations enrich them as well.¹¹ After the relations have been inherited, the users have the possibility to change them — delete some of them when not applicable or add new ones. The system also supports definitions of new relations and some (limited) inference with reverse and transitive operations. In addition, domain and range restrictions are taken into account.

The system is equipped with a ticket module. Thus, the workflow is organized in a more structured way with respect to the various expertise and responsibilities. Lemmas can be marked in a certain way that

¹¹Given that this transfer of relations is correct for the concepts represented by the synsets.

calls for the intervention of a more experienced user. Thus, a user could assign a ticket to a lemma in case of identification of some error, or suggest an edit of a particular synset. Each ticket contains two parts: a textual description of the problem, and a related topic (subject). There is a list of predefined topics for the tickets, created with respect to the workflow on BTB-WN, but other types might be added whenever needed. The list of current ticket subjects includes: *Edit synset*, *Missing sense*, *Wrong hypernym*, *Part of speech*, *General remark*, *Wrong equivalent*, *Discussion*, *Missing relation*, and *Link to Wikipedia*. In Fig. 2 an example is given. In this case the ticket relates to some missing senses for the selected lemma and provides suggestions for these senses. A more experienced user checks all the lemmas with such ticket subjects and approves the existing suggestions or adds the appropriate senses. The system allows for the users to search for lemmas with certain types of tickets. A result from such a search is given in left part of the figure. The search is for all lemmas that are marked with the subject for missing senses. The *Link to Wikipedia* subject is currently used for adding a Wikipedia link to the corresponding lemma in one of its senses. The tickets might be commented by other users (depending on their editing rights), then resolved and deleted. Information is also available about the author of the ticket, the date and the time when it was created, commented or deleted, who and when processed it.

Lemmas and synsets within the wordnets are in many-to-many relations which means that a lemma could belong to several synsets whereas a given synset could have more than one lemma. In both cases the lemmas and the synsets are not equal in their usages. Thus it is important to rank them with respect to their relevance. Currently, we rely only on the users' intuition for this step. Thus we aim at some initial lemma ordering for each synset with more than one lemma. In some cases the ordering is performed automatically. Such cases include, for example, the synsets for professions where the masculine form precedes the feminine one. Also in case of verbal synsets with imperfective and perfective aspect lemmas, the imperfective one comes first. The user interfaces for the ranking are given in Fig. 3. On the left side, the dialog for the arrangement of synsets is given. The user has a granted access to the categories and definitions of each synset. Thus, they can make an informed decision about the ranking. On the right, a similar dialog is presented where the user can check the synsets whose lemmas need to be ordered. In both cases the user could modify the ranking values. These rankings are used in different applications as Word Sense Disambiguation, selection of appropriate lemmas in text generation, etc. Needless to say, the manual ranking is subjective and thus not reliable enough, so in future more information about relations between lemmas and senses will be added to BTB-WN.

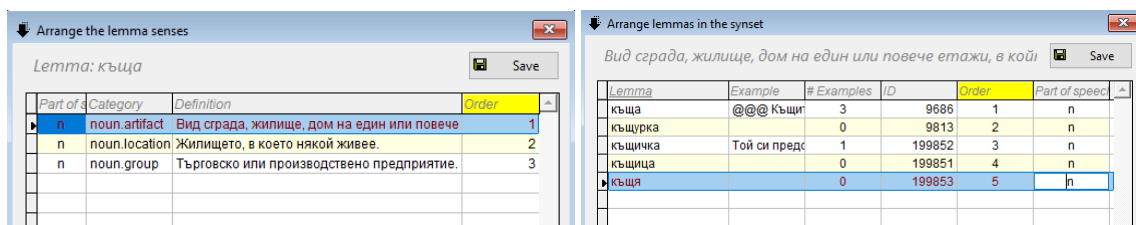


Figure 3: Arranging synsets in CLaDA-BG-Dict.

Our system provides more functionalities than the ones, introduced above. One of them is the access to corpora and machine readable dictionaries. It will be discussed in more detail in the next section. The remaining functionalities solve some smaller tasks like shortcuts assignments, procedures for automatic relation addition and similar.

3.2 Integration of Corpora and Other Dictionaries

In this section two of the main sources of information used by the lexicographers in their work — corpora and dictionaries — are discussed with respect to their integration within the CLaDA-BG-Dict System. The corpora are mainly used for searching examples for the various lemma senses and for finding new or missing senses. The dictionaries are valuable sources of different kinds of lexical knowledge.

The system allows for searching and adding example sentences directly (see Fig. 4). Users can provide any corpora relevant to their work. When selecting an example, the user can pin it to the corresponding

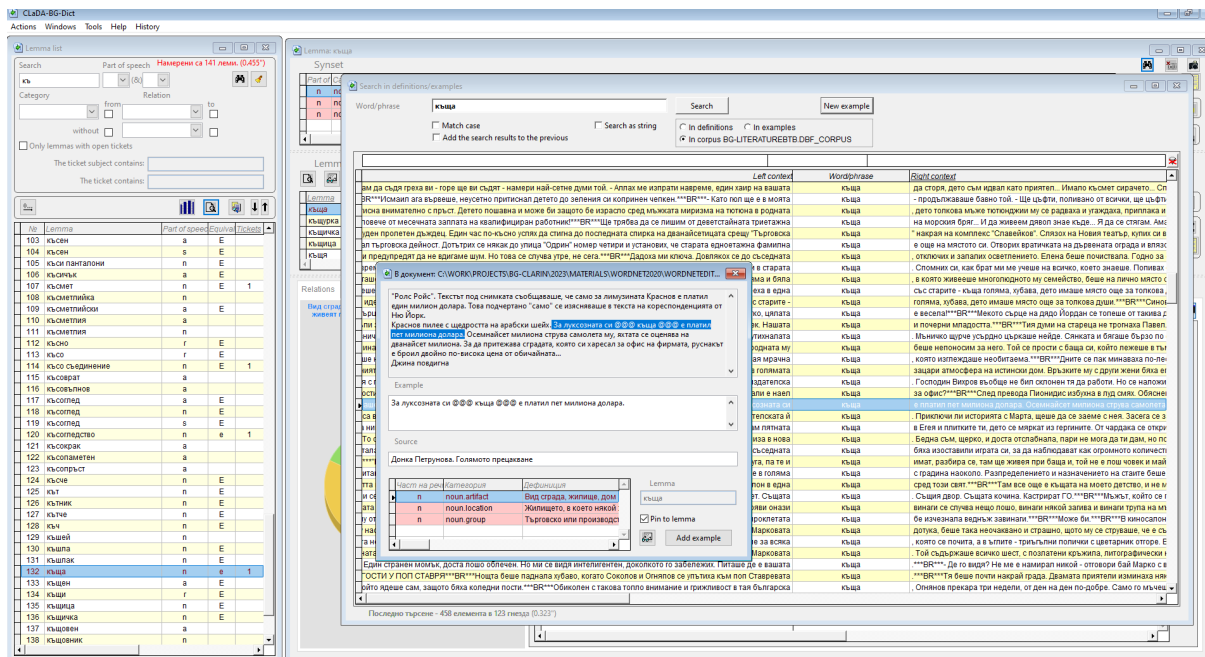


Figure 4: Search for examples in a corpus integrated in CLaDA-BG-Dict.

synset and lemma while the source of the example is copied automatically. The search could be performed by lemmas or strings. In future, the search engine will support also regular expressions. The system suggests that sentences are included as examples, but if the users decide otherwise, they can include also bigger contexts from source texts. The latter is especially important for the selection of examples that do not allow ambiguous interpretations. The assignment of examples to lemmas and synsets resembles the process of text annotation with senses from BTB-WN. For that reason, there are two special cases of corpora in the system that were already assigned to lemmas and synsets: the set of definitions and the set of examples. The user has the possibility to search in them and in this way to annotate all the words within definitions and examples. The intended usage of this option is the creation of corpora annotated with senses from BTB-WN, similarly to SEMCOR (Miller et al., 1993) and Gloss corpus (Rademaker et al., 2019). In addition to being used as sources for assigning examples to the existing synsets, the corpora are extensively used also for detecting new senses that are neither in the current version of BTB-WN, nor in the dictionaries. Currently we can only rely on the sorting functionality of the concordance with respect to the found items and their contexts. After having been sorted, the examples are checked one by one. In future we plan to use similarity measures over the context in order to cluster the concordance lines.

In addition to corpora we consider the access to existing machine readable dictionaries within the CLaDA-BG-Dict System as an important resource to be consulted during the creation of BTB-WN. The system provides access to four electronic dictionaries which are aligned through the lemmas they share. When an entry contains information about several lemmas, it is aligned to the other dictionaries through each of these lemmas. In this way the information for a given lemma is accessible through each of the lemmas in any of these dictionaries. Thus, users could observe all the information coming from the various dictionaries simultaneously. The four dictionaries integrated and actively used in the CLaDA-BG-Dict include: one explanatory dictionary — (Popov, 1994), one inflectional dictionary of Bulgarian — (Popov et al., 1998), (Popov et al., 2003), and two Bulgarian-to-English dictionaries — one freely available on the web and one based on our own Bulgarian vocabulary for bilingual dictionaries. The bilingual dictionaries are particularly useful for the selection of appropriate English equivalents from the EOW as well as for providing information about the number of senses for a given lemma. The

explanatory dictionary includes also information about idioms with the selected lemma, so they can be used as a source for creating synsets with multiword expressions. A problem which occurs during the integration of the various dictionaries is that there could be discrepancies among them of various kinds. For example, the dictionaries in the system provide as a rule different number of senses for a lemma. The reasons for this could be many but some of them are: some dictionaries include also archaic and dialectal senses and/or tend to distinguish among very similar senses, while others are more general and present only contemporary and/or gross-level senses. In addition, dictionaries are published at different times, so they show the senses typical for two or more different periods. Such contradictions between dictionaries are normal and should be expected, since all of them could follow different approaches and goals. An example of this issue from BTB-WN is the case with the noun *чета*: its most frequent sense found in all dictionaries is *a group of rebellions in liberation struggles*, but one of the integrated in the CLaDA-BG-Dict system dictionaries includes also an archaic sense *a group for common work* and a rare metaphorical meaning *gymnastics or other sport group*. The newer and more general dictionary in the system presents only the first most frequent sense nowadays. In such case, the lexicographers should follow the corresponding guidelines for the selection of senses.

Another challenge is related to the orthography – as mentioned above, the incorporated dictionaries in CLaDA-BG-Dict are published at different periods, so some of them are not complying to the current orthographic rules. At the same time, they all contain valuable information for senses, morphology, etc. and they are worth to be considered. The problem which however comes is that some lemmas would be flagged as not present in BTB-WN just because they are written differently.

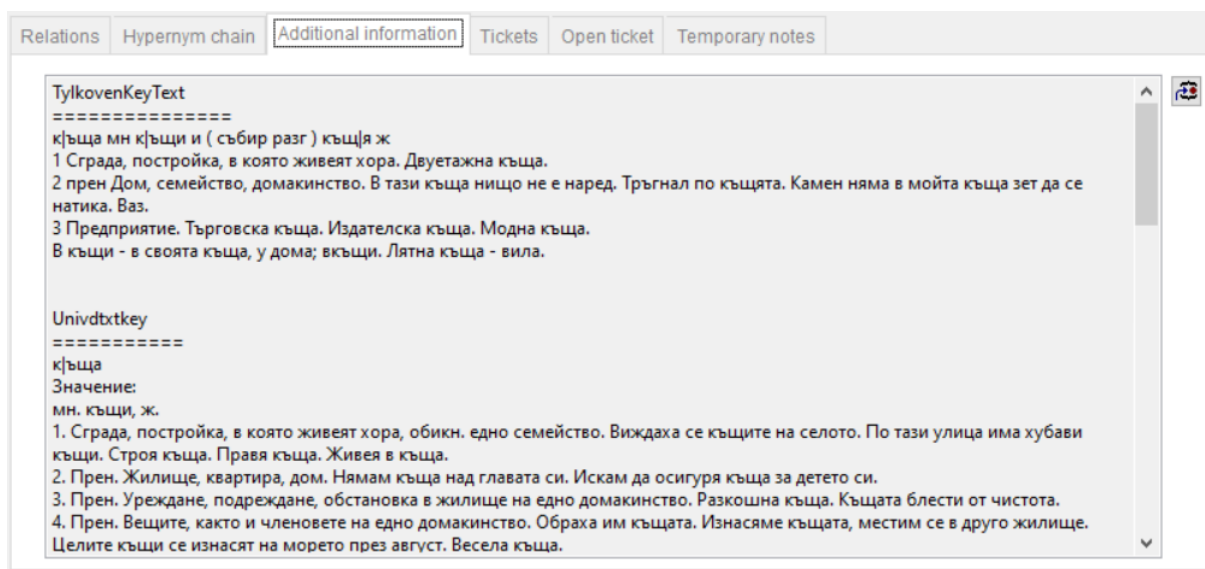


Figure 5: Access to dictionaries within CLaDA-BG-Dict via an editor form.

The access to the dictionaries is provided in two ways. The first one is from the editor form. In Fig. 5 information about the lemma *къща* in two explanatory dictionaries (which also provide idioms with the given lemma) is observed. The search is by the lemma associated to the editor form. The information from the dictionaries is presented in a tab from the Miscellaneous area. Such a look up in dictionaries is very convenient for quick checks of the various definitions, examples, English corresponding lemmas during the editing of existing synsets in BTB-WN. The second mechanism for a look up is independent from the editor form. There is an option for each dictionary to be opened in its own form, or another option when all dictionaries are opened in one form. The availability of these independent forms allows for searches for arbitrary lemmas, related lemmas, etc. A useful mechanism for access to the dictionaries is through the so-called Wordlist form. In this form a list of lemmas provided by the user is opened. The lemmas are checked whether there are already synsets related to them in the BTB-WN. If not, the search is done by pointing to a lemma in the Wordlist form and using a search shortcut which copies the lemma

to the dictionary form to perform the search.

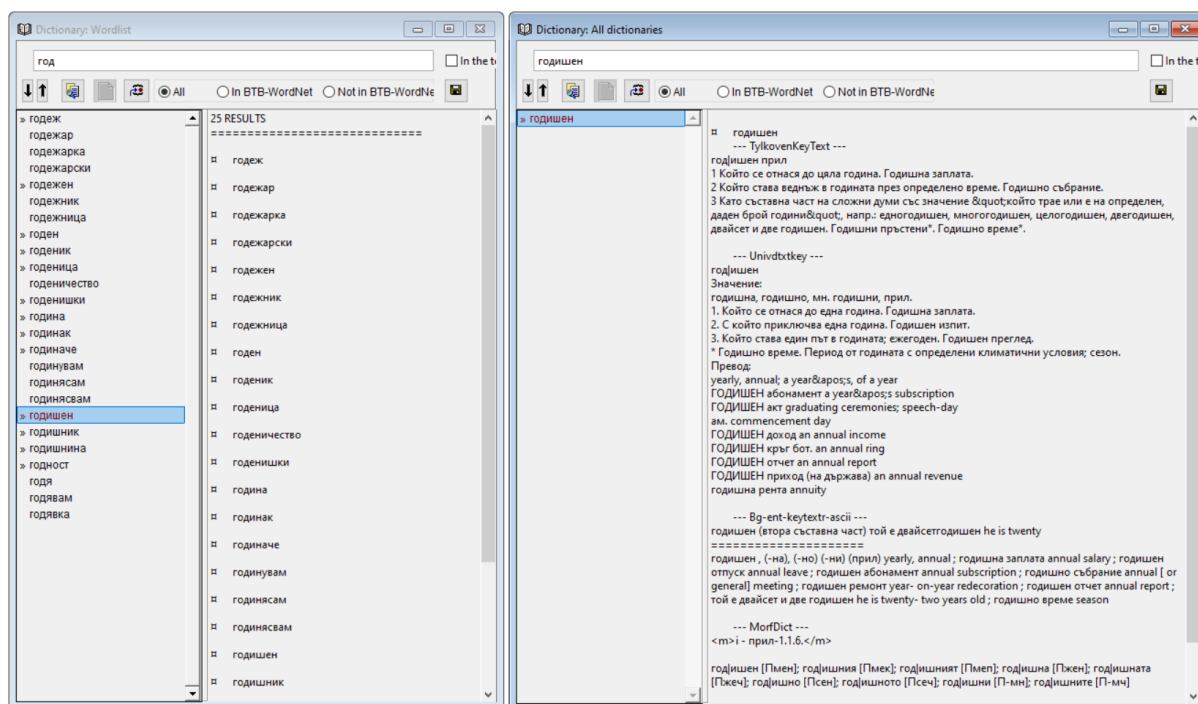


Figure 6: Look up in the dictionaries within CLaDA-BG-Dict in an independent form and with links to the Wordlist.

In Fig. 6 the two forms — All dictionary and Wordlist, are represented. The search in both forms is through regular expressions. In Wordlist we see all the words matching the search query. Selecting a lemma from the list and searching in dictionaries form would provide information from all lexicons.

With relying on different Wordlists, the users could examine some sets of lemmas in BTB-WN selected by certain criteria. In our work we consider several such sets like vocabularies corresponding to Bulgarian learners' levels like A1-A2, B1-B2, C; vocabulary for secondary school students, etc.

4 Conclusions and Future Work

As it was frequently stressed above, we aim to provide a system where the user will be able to exploit all the available dictionaries, corpora and services. At the same time (s)he will have the possibility to not only statically consult other dictionaries but also to search within corpora, make concordances, establish mappings, save and make publicly available the results of their work.

In our view the necessary minimum of functionalities of such a system would include: an editor of lexical entries that supports different structures of interrelated elements; access points to existing dictionaries and corpora; various types of searches and concordances, etc. BTB-WN has been fully developed in this system and serves as a connector to other dictionaries and corpora through its synset and lemma information.

In addition to uploading and making accessible new dictionaries, the system also supports mappings to Wikipedia via the inclusion of Wikipedia article URIs to the corresponding synsets. For now this operation works for equivalent concepts only, but more elaborated set of relations are necessary.¹²

Each of the included language resources inherits its structure defined in some standard (with some modification if necessary). For example, for a given lexicon included in the system the structure of the lexical entry will be presented in TEI Lex0¹³. Some other lexicon might be presented in Lemon or LMF. Thus, the user will be able to refer to the structure of the various lexicons, to extract parts from

¹²We plan to adopt an already existing schema like SKOS, LEMON, etc.

¹³<https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>

the lexicon entries, to combine elements from different types of lexicons. This will allow easy ways of reusing the available data. Another benefit of the system is that it can keep track on the provenance of the work threads. Currently the system keeps information about each editing operation, but we think more elaborate model is necessary.

When the new dictionary (lexicon) is shared within the system together with the relations to other resources, the result will be a valuable resource not only for supporting the future dictionary creation, but also for automatic processing.

CLaDA-BG-Dict is an editor, which could be used for both tasks – creating lexical databases like wordnets and compiling traditional types of dictionaries. But what is more – it provides possibilities of linking the available data in many ways depending on the goal. CLaDA-BG-Dict has already been successfully used for editing of more than 19 000 synsets that were created at earlier stages in an XML format, and for the addition of around 14 000 synsets together with appropriate examples. It thus provides quick access to various types of linguistic resources and information – dictionaries, corpora, concordance, etc. The resources are accessible in the system, so any kind of checks could be performed by the user in the same environment.

Our vision for future is to enhance replicability and re-usage of dictionary compilation for specific purposes as much as possible. In this way we believe that the work of dictionary creators and dictionary users will be facilitated and enriched.

Last but not least, in its beta-version now the system uses its own format for uploading corpora and other digitally-born or digitized dictionaries. However, it is planned that the system conforms to the common standards such as TEI, TEI LEX0, Lemon, etc. All the participating resources will be made available through the CLaDA-BG repository and dedicated web services.

Acknowledgements

This work was supported by *the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG*, Grant number DO01-301/17.12.21.

References

- Henrich, V. and Hinrichs, E. 2010. *GernEdit - The GermaNet Editing Tool*. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)
- Kallas, J. and Koeva, S. and Kosem, I. and Langemets, M. and Tiberius, C. 2019 D1.1. *Lexicographic practices in Europe: a survey of user needs*. ELEXIS - European Lexicographic Infrastructure. H2020 project.
- Měchura, M. 2017. *Introducing Lexonomy: an open-source dictionary writing and publishing system*. In: Proceedings of eLex 2017 conference. Leiden: Lexical Computing, 2017. p. 662–679.
- Miller, g. and Leacock, C. and Teng, R. and Bunker, R. 1993. *A semantic concordance*. Proceedings of the workshop on Human Language Technology. pages 303–308. Association for Computational Linguistics.
- Naskret, T. and Dziob, A. and Piasecki, M. and Saedi, Ch. and Branco, A. 2018. *WordnetLoom – a Multilingual Wordnet Editing System Focused on Graph-based Presentation*. Proceedings of the 9th Global Wordnet Conference. pp. 190–199
- Osenova, P., Simov, K. 2018. The data-driven Bulgarian WordNet: BTBWN. Cognitive Studies – Études cognitives, 2018(18) (2018) <https://doi.org/10.11649/cs.1713>
- Popov, D. 1994. *Bulgarian Explanatory Dictionary. Extended and Updated Edition. (in Bulgarian)* Nauka i izkustvo. Sofia, Bulgaria
- Popov, D. and Simov, K. and Vidinska, S. 1998. *Dictionary of Writing, Pronunciation and Punctuation of Bulgarian Language (in Bulgarian)* Atlantis KL, Sofia, Bulgaria
- Popov, D. and Simov, K. and Vidinska, S. and Osenova, P. 2003. *Spelling Dictionary of Bulgarian Language. (in Bulgarian)* Nauka i izkustvo. Sofia, Bulgaria

- Rademaker, A. and Cuconato, B. and Cid, A. and Tesseracto, A. and Andrade, H. 2019. *Completing the Princeton Annotated Gloss Corpus Project*. Proceedings of the 10th Global Wordnet Conference, pages 378–386, Wrocław, Poland. Global Wordnet Association.
- Simov, K. and Peev, Z. and Kouylekov, M. and Simov, A. and Dimitrov, M. and Kiryakov, A. 2001. *CLaRK-an XML-based system for corpora development*. Proceedings of the Corpus Linguistics 2001 Conference. pp. 558–560

Linguistic Autobiographies. Towards the Creation of a Multilingual Resource Family

Silvia Calamai

Università di Siena, Italy
silvia.calamai@unisi.it

Rosalba Nodari

Università di Siena, Italy
rosalba.nodari@unisi.it

Claudia Soria

CNR-ILC, Italy
claudia.soria@ilc.cnr.it

Alessandro Carlucci

Universitetet i Bergen, Norway
alessandro.carlucci@uib.no

Abstract

This paper describes a project aimed at creating a new resource family of multilingual and multimodal resources centered around the concept of a “Linguistics of the self”, i.e. personal reflections on the role of languages in shaping one’s identity. Language portrait silhouettes, drawing bilingualism, and linguistic autobiographies are different types of resources that share this common feature. We describe the resources and the criteria for their metadata annotation, focusing in particular on linguistic autobiographies, where the writer explicitly reflects on the relationship between him/herself and language. These genres are fruitfully used in different educational settings, and research has shown that they help to uncover the social, affective, and psychological dimensions of language learning. The potential of a multilingual and multimodal collection is discussed starting from data collected in Italy and Norway.

1 CLARIN Resource Families and Linguistic Diversity

The CLARIN Resource Family (Fišer et al. 2018) is a user-friendly overview per data type of available language resources in the CLARIN infrastructure aimed at the needs of researchers from digital humanities, social sciences, and human language technologies. Resource families are provided according to modality (spoken, multimodal, computer-mediated), genre (historical, academic, literary, newspaper, etc.), different languages, and intended use (reference, L2 learners). These groupings of corpora, lexical resources and other tools are meant to facilitate comparative research: for each resource family, a brief description is provided followed by a list of the resources belonging to the family, together with the most important metadata (name, size, annotation, license, language, description and availability). Thus, resource families provide a curated view of the available CLARIN resources. Over the years, this has proven to be a highly visible initiative appreciated by a broad spectrum of CLARIN users (Leonardič and Fišer 2020) and it therefore deserves to be maintained and enlarged. In this paper, we introduce a new genre that comprises various types of productions (texts and mixed text and drawings) sharing the common feature of containing personal reflections on the role of language in shaping one’s identity. We argue that a resource family devoted to this genre could be useful for both first and second language research and pedagogy, as well as for a better understanding of the linguistic landscape in European schools and universities, as it is shown in the following section.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>

Silvia Calamai, Rosalba Nodari, Claudia Soria and Alessandro Carlucci 2023. Linguistic Autobiographies. Towards the Creation of a Multilingual Resource Family. *Selected papers from the CLARIN Annual Conference 2022*. Ed. by Tomaž Erjavec and Maria Eskevich. Linköping Electronic Conference Proceedings 198, pp. 23–32. DOI: <https://doi.org/10.3384/ecp198003>

2 Self-Reflective Practices for Shaping Europe’s Multilingual Landscape

Pedagogical research has shed light on the importance of reflective practices for both teachers and students. As far as the former are concerned, reflective practices can help in reshaping their knowledge and their teaching practices (Farrell 2022). Students, too, can be encouraged to engage in self-reflection to help them focusing on the emotional and cognitive components of their attitudes toward the subjects they are studying. In language education there are several age-appropriate tasks that are typically employed in assisting students to reflect on their multilingual selves in order to develop meta-sociolinguistic competence. Among the most popular methods we can list: i) language portrait silhouettes; ii) drawing bilingualism; iii) linguistic autobiographies. Silhouettes and drawing bilingualism are discussed in §2.1, while a separate paragraph is devoted to linguistic autobiographies (§2.2).

2.1 Language Portrait Silhouette and Drawing Bilingualism

Language portrait silhouette (LPS) and drawing bilingualism (DB) are two multimodal self-reflective tasks that are suitable for early childhood and primary education because they do not require particular abilities in writing. However, LPS and DB can also be employed in other contexts such as secondary schools and universities, especially in multilingual classes where different levels of linguistic competence can be found. In LPS tasks, respondents are given a body silhouette (see fig. 1) and are asked to paint it, choosing different colors and body part for each language they want to mention (Busch 2010). They can be adopted also at a policy level: in the Autonomous multilingual Province of Bozen\Bolzano in South Tyrol, LPS is an accredited part of the European Portfolio of languages. Respondents are asked to fill the silhouette with different colors, using larger spots for the most frequently used languages, and to discuss their LPS with schoolmates (Provincia Autonoma di Bolzano 2004).



Figure 1. A template for LPS

Similarly, in the drawing bilingualism task (Favaro 2013), children can freely draw themselves without having a silhouette to fill in. Children have to imagine themselves as plurilingual individuals and choose which representation of themselves can better capture their plurilingual mind. Figure 2 shows a specimen collected in an Italian secondary school: according to the image, French is the language of the ‘head’ because it has been studied at school, while Italian is the language of the ‘heart’; German is the language of the ‘hand’ since the student has to struggle with its grammar and finally English is the language of the ‘legs’, because it opens up doors in space and time.

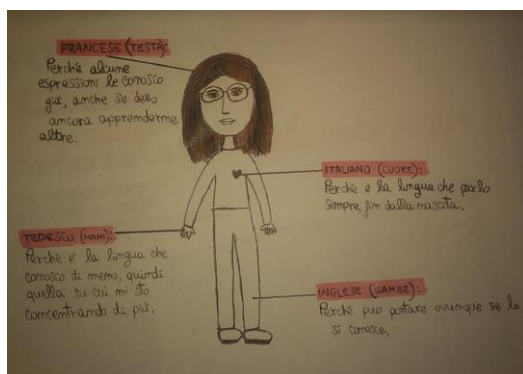


Figure 2. A drawing bilingualism example from the University of Siena corpus

These two tasks are commonly used across Europe (cf. for the Italian setting Manconi 2019; Biagioli 2021) as they do not require specific training; LPS can also easily be downloaded from different websites (see for example <https://cluss.unistrasi.it/public/articoli/157/Silhouette%20linguistica.pdf>). In these two tasks, the linguistic repertoire is truly interpreted as embodied cognition. With LPS and DB, researchers can look at linguistic repertoires from a multimodal perspective (Kress & van Leeuwen 2006): they can analyse the relation between language, colours and cultural tropes, the affiliation and emotional component of language learning, as well as the multimodal representation of language using different symbols. For teachers, LPS and DB could be used as an ice-breaker activity during the first days of school.

2.2 Linguistic Autobiographies

Unlike the two tasks explained in 2.1, linguistic autobiographies constitute a non-fictional genre where the writer explicitly reflects on the relationship between him/herself and language. Different labels have been used for describing this genre other than linguistic autobiographies, such as language memoirs (Pavlenko 2001, 2007; Kramersch 2004) or language stories (Gohard-Radenkovic, Rachedi 2009). In this self-reflective writing practice, language becomes the overarching organising principle for re-tracing salient moments in the writer's life. The idea behind this genre is that the acquisition and the interaction of different languages can be seen as the acquisition of selfhood (Ramsdell 2004). Linguistic autobiographies can be considered as both a research and a pedagogical tool. They are used by professors and teachers in secondary schools and university classrooms to help students in developing their metalinguistic and metapragmatic abilities; within superdiverse multilingual classrooms, linguistic autobiographies allow students to narrate their multilingual selves and they help make their languages more visible. Linguistic autobiographies can also help linguists in gaining access to language ideologies and attitudes towards language varieties: in particular, these narratives can help understand how ideologies about languages can have an impact on the linguistic behaviour of speakers.

Linguistic autobiographies are highly versatile in that they can easily be collected without requiring specific skills or academic knowledge. Several templates are available to facilitate the production of linguistic autobiographies, which offer some suggestions to think about languages and self-reflect on key points of the writers' own life (cf. Canobbio 2006; D'Agostino 2007; Luppi, Thüne 2022)¹.

For example, one possible template requires mentioning:

- 1) Family members and personal data (place of birth, eventual relocations, etc.);
- 2) Family linguistic background: L1 of the grandparents, L1 of the parents;
- 3) Family linguistic situation: parents and grandparents' linguistic preferences (they speak which language to whom and when); which languages are used for ordinary communication between family members (with children, with the rest of the family); family choices in linguistic education (which

¹ The Council of Europe provides its own template for compiling language biographies of users' plurilingual profiles. The scheme can be found here: <https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016804932c5>)

language is taught to children); which language is spoken at home; which other varieties are used at home, and for what purposes (communicative, expressive, affective needs, identity stances, etc.);

4) Family and school attitudes and behaviours: repression of non-standard varieties; are any specific varieties preferred to other varieties? Are there any disfavoured languages? Are there any forbidden languages used in secret between friends or family members?

5) Meeting with linguistic diversity (holiday trips to different regions, community of practices, peer groups, school environment, extended family etc.). Formal and informal language learning (foreign language schools, friends from abroad etc.); are different languages used with different groups of people? Are non-standard varieties used for performing specific identities? Etc.

6) Personal evaluation of language learning in and out of school; ability to perceive different linguistic varieties, social evaluation of accents, stigma and stereotypes towards accents, varieties, languages, etc.

Here is an instance of a linguistic autobiography produced by a second-generation Italian student of Chinese origin, followed by its translation into English.

Io ho iniziato a studiare l'italiano quando andavo all'asilo e l'ho imparato insieme al cinese.

Quando andavo all'asilo soprattutto io parlavo italiano perché c'erano solo 2 o 3 compagni cinesi. A casa io parlo cinese con i miei genitori e alcune volte italiano o inglese con mia sorella, i miei genitori parlano sempre il dialetto e vuole anch' io parlare il dialetto a casa e mi dicono non devi mai dimenticare il dialetto perché questo è il nostro radice. Io vorrei migliorare il mio italiano perché questa lingua l'ho iniziato a studiare a l'asilo e per un altro motivo: io abito in Italia e userò più italiano di cinese. Io vorrei migliorare francese e tedesco più d'italiano perché queste due lingue forse le userò per lavorare².

I started studying Italian when I was in kindergarten and I learned it together with Chinese.

In kindergarten I used to speak Italian because there were only 2 or 3 Chinese classmates. At home I speak Chinese with my parents and sometimes Italian or English with my sister, my parents always speak the dialect and I want to speak the dialect too at home and they tell me you must never forget the dialect because this is our root. I would like to improve my Italian because I started studying this language in kindergarten and for another reason: I live in Italy and I will use more Italian than Chinese. I would like to improve French and German more than Italian because I will perhaps use these two languages for work.

The template (and the terminology used) can be adapted depending on the age, educational background and other characteristics of those involved. In any case, linguistic autobiographies are deeply personal texts that outline the writers' own thoughts and feelings, giving them the possibility for spontaneous expression.

2.3 The Societal Potential of Self-Reflective Practices in Education

All these tasks are fruitfully used in different educational settings, and research has shown that these tools help to uncover the social, affective, and psychological dimensions of language learning (Franceschini, Miecnikowski 2004; Groppaldi 2010; Cavagnoli 2020; Salvadori, Blondeau, Polimeni 2020). Language biographies are also an important element of the European Language Portfolio, as they encourage learners to write information on linguistic and cultural experiences gained in and outside formal educational contexts (Council of Europe 2001). They permit students to develop an awareness of cultural and linguistic diversity, and to learn about the social value of languages. In superdiverse settings, linguistic autobiographies help students in understanding mechanisms of stereotyping and linguistic discrimination and are considered an empowering tool. Teachers can also gain access to the students' linguistic learning process, and they can discover students' learning practices and reasons for studying languages, as well as their needs and expectations. For policy makers and stakeholders, self-reflective practices can help in understanding students' motivation for language learning, thus allow-

² The linguistic autobiography reported in the paper is directly quoted from the original text, without any editing from the authors.

ing them to develop specific school curricula addressing students' communicative needs. For those who are interested in the sociology of languages, these data can also provide unique information about informal language-learning opportunities and the different values that speakers attach to different types of multilingualism.

3 Towards a CLARIN Resource Family for Self-Reflective Practices

Language portrait silhouettes, drawing bilingualism, and linguistic autobiographies constitute different genres of great scholarly significance and with high potential of impacting education. The genres above can all be considered self-reflective practices, with language(s) being the overarching characterizing principle. For this reason, we believe that a multilingual and multimodal resource family that collects LPS, DB and linguistic autobiographies produced by L1 and L2 speakers of different languages in different countries is an important contribution to the scholarly community gathered around CLARIN. This new resource family will thus deal with peculiar genres with specific formal and content characteristics, and it is also open to hosting other genres and text types, such as oral interviews and personal narratives. From our point of view, it is desirable that all textual and non-textual products sharing the trait of personal metalinguistic reflection can be collected, and that their collection is carried out in ways that allow scholars to easily compare them and practitioners to use them immediately. The new resource family, called "Linguistics of the Self", will thus offer the possibility for scholars to access all these kinds of data, and to contribute with their own data. In this section we report about the principles and criteria that we will follow in building the resources and making them available as a CLARIN resource family. We will provide examples from the corpus of linguistic autobiographies, which so far represents the largest component of the family.

3.1 Adding Linguistic Autobiographies to CLARIN

The University of Siena corpus of linguistic autobiographies will be the core of this new multilingual and multimodal resource family. At present, this resource consists of about 300 linguistic autobiographies collected during university courses in linguistics (educational linguistics and sociolinguistics)³, about 50 autobiographies written by secondary school students, and about 40 produced by secondary school teachers. An example is available in Figure 3. In addition to the core, the resource will also include autobiographies collected in Norway. A small pilot collection has been carried out at the University of Bergen, in Norway, and has produced texts in different languages, including English and Spanish (see Fig. 4). The collected texts will be digitized, and data processing will be carried out for anonymization, automatic linguistic analysis with the Profiling-UD tool (see Brunato et al. 2020) and metadata description, on the basis of a shared procedure. A strong requirement that constrains how the corpus is going to be structured is that it should be easy to update, so that the corpus can be integrated with new linguistic autobiographies. Indeed, linguistic autobiographies, as well as LPS and DB, are commonly produced and collected by others, in addition to the original creators of the resource. Possible potential contributors are schoolteachers who collect this type of data according to the procedure that we suggest, but also other university professors. Linguistic autobiographies are currently being collected as part of teaching activities at several Italian universities (University of Pavia, Macerata, Roma Tre, among others), especially in educational linguistics and sociolinguistics courses. Even if this resource is not particularly large in number, compared to other CLARIN resources, one of its main strengths is its transversality, in that it can be used in any higher education degree. This will allow the resource to grow steadily each year.

In this respect, we have identified the structure of the ParlaMint project (Erjavec et al. 2022), where each file of the corpus represents a single transcript session, as the most appropriate for our needs.

³ At the University of Siena, approximately 60 autobiographies per academic year are collected. In particular, the collection is an integral part of the course in Sociolinguistics: autobiographies are collected at the beginning and the end of the course and function as a tool both to introduce some key notions from sociolinguistics and to allow individual reflection by students at the end of the course, once analytical skills and terminology have been acquired.

This will allow the corpus to grow constantly, with new autobiographies being periodically added to the original collection.

SEMPRE RIMASTA A VIVERE IN QUELLA ZONA
 GROSATA SOLO RESIDENZA TRA DUE COMUNI
 LIMINAPOLI -
 - PADRE NATO NELLO STESSO COMUNE
 - MADRE NATA IN ALTRO COMUNE, MA A SOLI 15 KM.
 ANCHE LOO ~~GENITORE~~ VISSUTO NELLA STESSA VALLATA
 IL CASENTINO (ALTO CASENTINO)
 - ~~IO~~ NONNE VENIVANO DALLA STESSA AREA GEOGRAFICA
 MA IN GIOVENTÙ SI ERANO TRASFERITI PER
 LAVORO A FIRENZE O PROVINCIA
 - HO VISSUTO CON UNA NONNA CHE PARLAVA "OSCANO"
~~OSCANO~~ CON UNA CADENZA INCARCATI DI
 DIALETTO NEMPOLESE / COME ANA /
 * LEL USAVA PAROLE DIVERSE DAGLI ALTRI COMPONENTI
 DELLA FAMIGLIA (E NE ANDAVA ORGOGLIOSA).
 - HO STUDIATO E HO SONO LAUREATA IN LINGUE STRANIERE
 (INGLESE - SPAGNOLO) HO SOGGIORNATO ALL'ESTERO
 - QUANDO ~~ARRIVAI~~ AD ABBEZZO A SCUOLA GLI AMICI MI
 CHIEDEVANO SE VENIVO DA FIRENZE "PARLI FERENTINO
 ?? A ME NON SEMBRAVA AFFATTO !!
 - QUANDO OSPITAVO AMICHE STRANIERE A CASA
 MIA NONNA PARLAVA IN ITALIANO, MA ALZANDO IL
 TONO DICEVA "COME SE FOSSEVO STATE SORDE" !!

Figure 3. A hand-written linguistic autobiography from the University of Siena corpus

Autobiografía lingüística

En mis 20 años de vida he aprendido más o menos los siguientes idiomas: el inglés casi como si mi existencia dependiera de ello, el alemán por obligación, el italiano por gusto y el español, mi lengua materna, quizás nunca lo termine de aprender.

Desde el primer año de colegio empecé a escuchar el inglés. En mi casa nadie hablaba un idioma diferente al español así que aprendí la pronunciación inventada por mis papás tratando de leer el "Workbook", más adelante mis profesores de bachillerato sufrieron para corregir esa pronunciación a la que estaba acostumbrada. Mis papás se niegan completamente a que viva en Colombia e hicieron todo lo posible para que yo aprendiera inglés, por eso siempre estuve en colegios bilingües donde enseñaban inglés británico e insistían en que escucháramos música, viéramos películas y leyéramos todo en inglés. Dado que los libros que me gustaban se publicaban primero en inglés, aprovechaba la excusa de practicar para que me los compraran o para que me dejaran pasar más tiempo viendo televisión, y llegó un punto en el que mi entorno estaba tan plagado del idioma que se me hacía raro cuando alguien me contaba que no lo hablaba.

Al entrar a la universidad no tuve que cursar las clases obligatorias de inglés porque ya tenía un buen nivel, para abrir la posibilidad de irme a otro país empecé a estudiar alemán. Me ilusionaba la idea de leer los textos teóricos de mi carrera en su idioma original y entender canciones, pero la emoción se fue tan rápido como llegó, las clases pasaron a la virtualidad por la pandemia y me costaba mucho concentrarme o siquiera entusiasmarme por ver la ventana de zoom presentar un libro de ejercicios mientras la voz del profesor iba y venía. Tan pronto como se acabó el curso dejé de practicar y olvidé casi completamente todo lo aprendido, cuando se dieron cuenta mis papás se enojaron diciendo que era desperdiciar una oportunidad muy grande y que debería seguir estudiando alemán a pesar de que no me sintiera verdaderamente motivada ni quisiera hacerlo.

Aprendí italiano por una canción que me gustaba mucho y más tarde descubrí que la letra está en napolitano. Al inicio entraba a las clases con miedo de que se repitiera la experiencia del alemán, de hecho ni siquiera le conté a mis papás que me inscribí al curso. Sin embargo fue totalmente diferente a lo que esperaba y mis dudas se transformaron en entusiasmo y pasión por el idioma, gracias a la clase de italiano conocí personas maravillosas con las que me encanta hablar y practicar a pesar de que hace tiempo concluimos el curso. Además, la profesora me hizo entender que un idioma no es la traducción del otro, y nos pedía que pensáramos directamente en italiano y nos expresáramos con las herramientas que teníamos aunque al principio fuera difícil. Así fue como sentí que conecté más con el idioma y me interesé por la historia y el entorno que lo hicieron ser como es, incluso empecé a interesarme esos aspectos de mi propio idioma.

El español ha sido una cuestión compleja porque he vivido en tantas ciudades diferentes de Colombia que mi acento es irreconocible, aunque preferiría que se notara el acento de la ciudad donde nací inconscientemente adopté las formas de hablar del sitio donde estoy, eso le encanta a mi mamá porque dice que en mi tierra la gente es muy corroncha y no saben hablar. Ella odia todo lo que no sea español académico, al punto de que considera a algunos géneros musicales, como el vallenato, inferiores. Después de varias peleas con ella por mi interés en la tradición cantada del español, pude disfrutar de esa historia tanto amo y que hace parte de mí. Ahora, junto a mi deseo de viajar y vivir en diferentes países, está también el deseo de atesorar y llevar conmigo la belleza de mi idioma.

Figure 4. A linguistic autobiography collected at the University of Bergen

This new CLARIN resource is designed to be comparable with other types of resources, such as LPS and DB, that share some of the features of linguistic autobiographies. Additionally, linguistic autobiographies could share some features with already existing or future resources, such as focus groups on language attitudes, qualitative interviews and oral narratives in which the interviewee describes his/her relationship with his/her mother tongue(s), language acquisition, etc.

This approach derives from the fact of having devised our resource as a member of a resource family from the outset, and it has important consequences for the choice of metadata (see §3.2). Furthermore, it is important to us that the new resource can easily be compared with similar resources that are

already part of CLARIN collections. For example, by searching the corpus according to speakers' L1, it will be possible to compare the biographies of speakers with the same L1 across different countries, educational settings, etc. This possibility may be used for various purposes, e.g. for a comparative analysis of the effects of language contact in different settings and with different L2s. Therefore, we foresee that rich metadata will be added to the corpus of linguistic autobiographies: all the valuable sociodemographic data, the language used and all the languages and varieties mentioned in the autobiographies will have to be made explicit. This will greatly enhance the usability and comparability of the resource (see below).

3.2 Criteria for Metadata Selection for a Multilingual and Multimodal Resource Family on the “Linguistics of the Self”

The choice of metadata is a crucial issue in building resource families (Leonardič and Fišer 2020), and it is therefore of paramount importance. When a resource family is created from already available resources, it is often the case that the curated resources have different depths and breadths of metadata description, which in turn has consequences for their final usability. In our case, where the individual resources are going to be created and described having in mind their role as members of a wider resource family, it is of utmost importance that all the different genres are described in such a way that their collection under a resource family is straightforward and allows for their maximum comparability. Here we introduce some of the criteria that will guide us in the selection and adoption of metadata for the resources composing the family.

We believe that a metadata description that is oriented to making a resource fit in a resource family must satisfy two interacting main requirements: a) exhaustiveness and b) comparability.

Exhaustive metadata description is important not only for the sake of describing any given individual resource accurately, but also in order to maximise its traceability and the possibility for it to be discovered and included in future collections. All aspects belonging to the “linguistics of the self” will need to be highlighted accordingly: the self-descriptive aspect, the modality of the task (spontaneous drawing, written autobiography, oral interview, etc.), all the languages that are mentioned and the language in which a given text is produced, whether the language is an L1 or L2 for the author, in addition to more customary socio-demographic factors such as sex, age and provenance. Metadata description will also take into account other important aspects, namely school grade (resources collected in primary or secondary schools, or universities) and metalinguistic competence (autobiographies collected before and after linguistic courses).

Making this information explicit will make it possible, for instance, for a scholar interested in studying the attitudes of students towards English as a second language to select all autobiographies written in different languages and mentioning English as a second language.

A TEI Header will be developed to encode this information. Each autobiography will then be annotated according to the TEI header scheme. The TEI Header scheme will be made available to download, in order to facilitate the annotation of similar resources already owned by other scholars.

In order to ensure the highest possible degree of comparability with other resources, either already present in the CLARIN ecosystem or to be added in the future, metadata description for the new resources will have to take into account the metadata sets used for describing resources that are partially overlapping for content and/or genre with linguistic autobiographies. From a first analysis of the VLO we have identified oral interviews, general autobiographies and personal narratives as the most similar genres already represented in CLARIN collections. The metadata used for describing these resources will be studied and analyzed to identify which elements of their description can be included in metadata description for linguistic autobiographies and linguistic silhouettes.

Since these are very peculiar genres not currently available in CLARIN, none of the available profiles for describing the resource is entirely suitable. The VLO repository, for instance, already offers a selection of linguistic autobiographies collected in two (non-exclusively) Italian-speaking settings, namely Language Biographies from South Tyrol and from Basel. However, these two corpora consist of audio interviews, together with their transcriptions. Written linguistic autobiographies, on the other hand, show some peculiar features such as a) written modality vs. mainly oral modality, and b) their strong emphasis on the linguistic component: language is the key around which the narrative is built

and articulated, and the recollection of one's life follows a linguistic path, while, in general, oral narratives focus on the main events in the lives of speakers.

3.3 Working Agenda

Available CMDI profiles will be studied and, if necessary, an ad hoc profile will be created to adequately describe the new resources. In addition to the metadata definition, legal and ethical issues will be addressed for personal data protection, in order to clarify i) the legal basis for future fieldwork collection, especially in the case of children/minors; ii) the procedure for obtaining consent for texts already collected – also before the introduction of GDPR; iii) anonymization strategies, especially in cases where explicit consent is absent.

In addition to the resource, support material for the collection of new resources will then be made available: three scripts (in English, Italian, and Norwegian) to guide the researcher/teacher in the elicitation of the texts (linguistic autobiographies, LPS, DB); guidelines for anonymizing the texts; and a template for their representation in TEI format. The script for the collection of autobiographies will only contain minimal information, such as the one mentioned in 2.2. For LPS, the silhouette will be provided, together with general instructions in three languages. The proposed family will thus provide a consistent and shared framework for the collection and curation of texts and multimodal products that are remarkably peculiar and therefore highly subject to the risks incurred by any data-scarce collection, such as inconsistency of structure, format, and level of analysis. The common template for metadata description of the different resources belonging to the family and the protocol for elicitation of the interviews will prove useful for ensuring coherence and comparability across different initiatives and will greatly help researchers and teachers by providing them with an established format.

Alongside this investigation, we will implement specific activities to promote linguistic autobiographies as an educational tool, also in order to discover any uses already in place but unknown to us. We are confident that in some of the countries involved in the CLARIN network, linguistic autobiographies are already used in school and university settings. In this respect, ILC-CNR and the University of Siena have now the advantage of being involved in the CIRCE Erasmus+ project (2023-25) aimed at studying accent discrimination in education. The school environment is a hotspot for investigating this issue: students are exposed to different accents and form and reinforce their attitudes and beliefs towards them also on the basis of peer pressure. Teachers are also confronted daily with regional and non-native accents of the national language, and may unconsciously succumb to prejudice and negative evaluations of non-standard varieties. In the project, a specific task will be devoted to training school teachers in using linguistic autobiographies in their classes. These autobiographies will then be FAIRified and will be added to the CLARIN resource family. This will permit to collect at least 200 multilingual autobiographies in different school settings and in different countries (in a digital format).

We thus wish to help uncover any already existing material and encourage the production of new material. A multilingual collection of such written material will indeed offer an invaluable picture from several perspectives. Firstly, it can be used as teaching material, from school classes of any grade to university courses, in order to raise awareness of heritage languages, accentism and glottophobia. Secondly, it can help teachers to better understand the most used and known languages in their classrooms. And thirdly, this collection represents a useful tool to verify – among pupils, students and teachers – the pervasiveness of the concept of linguistic error and deviation in describing linguistic repertoires.

Comparable corpora of linguistic autobiographies will also provide valuable quantitative and qualitative data to researchers interested in a variety of topics – such as language attitudes, language and migration, multilingualism and language contact. Finally, this new resource can help policymakers in designing linguistic policies that are more consistent with the different linguistic landscapes existing in different European schools and universities.

Acknowledgments

For Siena University and CNR authors, the project described in the text is partially co-funded by the European Union (CIRCE project (Erasmus+ AGREEMENT NUMBER: 2022-1-IT02-KA220-SCH-000087602).

References

- Biagioli, R. 2021. Promuovere un'educazione linguistica interculturale. *Educazione interculturale*, 19(2): 33-45.
- Brunato D., Cimino A., Dell'Orletta F., Montemagni S., Venturi G. 2020. Profiling-UD: a Tool for Linguistic Profiling of Texts. In *Proceedings of 12th Edition of International Conference on Language Resources and Evaluation (LREC 2020)*, 11-16 May, 2020, Marseille, France.
- Busch, B. 2010. School language profiles. Valorizing linguistic resources in heteroglossic situations in South Africa. *Language and education* 24(4): 283-294.
- Canobbio, S. 2006. Dialecto dei giovani e politiche linguistiche delle famiglie, appunti dal Piemonte. In: Marcato, G.: *Giovani, lingue e dialetti, Atti del Convegno*, Sappada - Plodn, 29 June - 3 July 2005. Unipress, Padova: 239-244.
- Cavagnoli, S. 2020. Diventare insegnanti: l'importanza della riflessione di futuri/e docenti sul proprio percorso di apprendimento linguistico. *Italiano LinguaDue* 12(2): 146-160.
- Council of Europe 2001. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press, Cambridge.
- D'Agostino, M. 2007. *Sociolinguistica dell'Italia contemporanea*. Il Mulino, Bologna.
- Erjavec, T., Ogrodniczuk, M., Osenova, P. et al. 2022. The ParlaMint corpora of parliamentary proceedings. *Lang Resources & Evaluation*. <https://doi.org/10.1007/s10579-021-09574-0>
- Farrell, T. S. 2022. *Reflective practice in language teaching*. Cambridge University Press, Cambridge.
- Favaro, G. 2013. Il bilinguismo disegnato. *Italiano LinguaDue*, 5(1): 114-127.
- Fišer, D., Lenardič, J., and T. Erjavec. 2018. CLARIN's Key Resource Families. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*: 1320–1325.
- Franceschini, R., and Mieczkowski, J. (eds) 2004. *Leben mit mehreren Sprachen. Vivre avec plusieurs langues. Sprachbiographien. Biographies langagières*. Lang, Bern.
- Gohard-Radenkovic, A., and Rachedi, L. 2009. *Récits de vie, récits de langues et mobilités*. L'Harmattan. Paris.
- Groppaldi, A. 2010. L'autobiografia linguistica: strumento per una moderna didattica dell'italiano L2-LS. *Italiano LinguaDue*, 2(1): 89-103.
- Kramsch, C. 2004. The multilingual experience: insights from language memoirs. *TRANSIT*, 1(1). <http://dx.doi.org/10.5070/T711009691>
- Kress, G., & van Leeuwen, T. 2006. *Reading images: The grammar of visual design*. Routledge, London.
- Leonardič, J. and D. Fišer. 2020. Extending the CLARIN Resource and Tool Families. In: Navarretta, C. & Eskevich, M.: *Proceedings of CLARIN Annual Conference 2020*, 05-07 October 2020, Virtual Edition: 1–5.
- Luppi, R, and Thüne, E.-M. (eds) 2022. *Biografie linguistiche. Esempi di linguistica applicata*, Centro di Studi Linguistico-Culturali (CeSLiC), Bologna.
- Manconi, T. 2019. La “silhouette des langues”: dessiner et représenter les identités plurielles. *Annali online della Didattica e della Formazione Docente*, 11(17): 107-123.
- Pavlenko, A. 2001. Language learning memoirs as gendered genre. *Applied Linguistics*, 22(2): 213-240.
- Pavlenko, A. 2007. Autobiographic narratives as data in applied linguistics. *Applied Linguistics* 28(2): 163–188.
- Provincia autonoma di Bolzano 2004. *Portfolio europeo delle lingue, per alunne e alunni dai 9 agli 11 anni*. Direzione Istruzione e Formazione tedesca, Bolzano.
- Ramsdell, L. 2004. Language and Identity Politics: The Linguistic Autobiographies of Latinos in the United States. *Journal of Modern Literature* 28(1): 166-176.

Salvadori, E.; Blondeau, N.; Polimeni, G. (eds.) 2020. Lingue maestre. autobiografia linguistica e autoformazione dei docenti di L1 e L2. *Italiano LinguaDue*, 12(2).

The Pipeline for Publishing Resources in the Language Bank of Finland

Ute Dieckmann
ute.dieckmann@helsinki.fi

Mietta Lennes
mietta.lennes@helsinki.fi

Jussi Piitulainen
jussi.piitulainen@helsinki.fi

Jyrki Niemi
jyrki.niemi@helsinki.fi

Erik Axelson
erik.axelson@helsinki.fi

Tommi Jauhiainen
tommi.jauhiainen@helsinki.fi

Krister Lindén
krister.linden@helsinki.fi

Department of Digital Humanities
University of Helsinki, Finland

Abstract

We present the process of publishing resources in Kielipankki, the Language Bank of Finland. Our pipeline includes all the steps that are needed to publish a resource: from finding and receiving the original data until making the data available via different platforms, e.g., the Korp concordance tool or the download service. Our goal is to standardize the publishing process by creating an ordered checklist of tasks with the corresponding documentation and by developing conversion scripts and processing tools that can be shared and applied on different resources.

1 Introduction

The Language Bank of Finland (Kielipankki, “The Language Bank”) is a collection of services for researchers using language resources in digital humanities and social sciences. The Language Bank is coordinated by FIN-CLARIN, a Finnish consortium of universities and research organizations. The general goal of the Language Bank is to make corpora and related tools available to users. Various types of resources can be deposited in the Language Bank, including text and speech corpora, lexicons and terminologies, and many kinds of data sets produced by research projects.

The Language Bank supports public, academic as well as restricted license categories and offers multiple services for providing access to different resource variants. Since the publication framework is complex and not yet sufficiently automatic, depositors cannot upload their resources to the Language Bank and publish them there directly. However, the Language Bank helps and supports the depositors in clearing the licenses and in converting, annotating and describing their data. Thus, unlike other CLARIN centres, the Language Bank participates to some extent in most of the steps in the process where a researcher or a research group deposits a resource with the Language Bank for redistribution. There are both advantages and disadvantages to this approach, which will be discussed in this article.

Ideally, all resources published via CLARIN services should meet the FAIR standards: they should be findable, accessible, interoperable as well as reusable¹. By creating a shared and well-documented workflow and by using common tools, we aim to ensure that all resources and their future versions are processed, published and maintained in a consistent, transparent and interoperable way.

2 The Publication Framework

An overview of the publication framework of the Language Bank is shown in figure 1. The process is started by entering the new resource to the publishing pipeline (cf., the left column of the figure). The

This work is licensed under a Creative Commons Attribution 4.0 International License. License details:
<http://creativecommons.org/licenses/by/4.0/>

¹ <https://www.clarin.eu/fair>

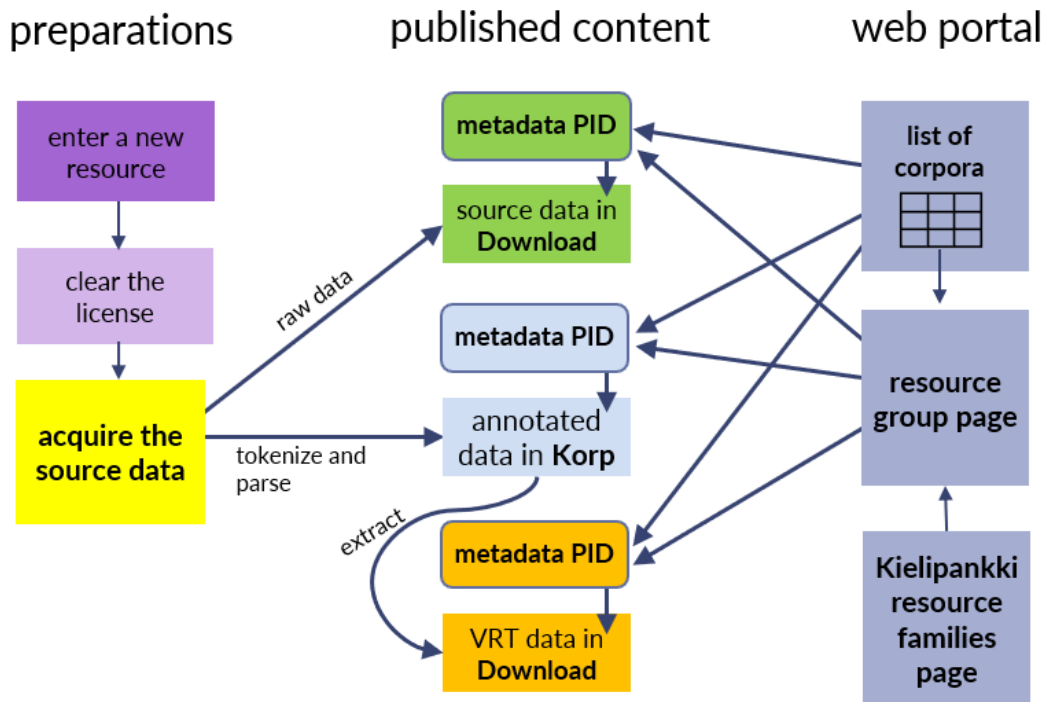


Figure 1. The structure of the resource publication framework

data is then described and prepared for the different means and formats of publication (the middle column). In order to make it easy for users to locate and to use the corpora they need, various pages are created on the website of the Language Bank (“the portal”) with additional information about the available resources (the right side of the figure).

2.1 Means of Publication

The resources published in the Language Bank of Finland may be available via the online concordance tool Korp, developed by Språkbanken, the Swedish Language Bank (see Borin et al., 2012), and adapted for the Language Bank of Finland². Korp is a web-based tool that allows the user to search for keywords and more complex constructs in text corpora that are typically enriched with grammatical and other types of annotations. In Korp, it is possible to generate concordances, to compute statistics based on various attributes in the corpora, and to download the results. Korp is well supported and it is used in several CLARIN centres.

Resources can also be downloaded via the download service³ of the Language Bank. In the download service, we usually provide the original source data as well as the converted data in VRT (*VeRticalized Text*) format as extracted from Korp. Several variants can be offered of the same resource. For the users’ convenience, copies of selected versions of the downloadable corpora are also accessible in the computing environment at CSC – IT Center for Science⁴. This makes it easier for users to process the content since they do not have to download and unzip large packages, and the required software can be made readily available on the server.

Lexical resources can be made available via Sanat⁵, a WikiMedia-based platform. However, Sanat is currently not part of our official pipeline, since the content is maintained in a community-driven fashion by individual research groups.

The Korp version in the Language Bank supports video and audio playback to a limited extent. In case a speech corpus includes transcripts of the original recordings, the transcripts can be used to create

² <https://www.kielipankki.fi/korp/>

³ <https://www.kielipankki.fi/download/>

⁴ <https://www.csc.fi>

⁵ <https://sanat.csc.fi>

a text version of the corpus. The texts can be equipped with links to the corresponding media files, helping the user to locate the original recordings when needed. In case the transcribed text has been manually or automatically aligned with the original recordings, the words and sentences in the transcripts can be annotated with time stamps. The timing information then allows the Korp interface to play the corresponding portion of the media to the user when requested. Korp does not include features for analysing speech signals or video content. However, Korp can be used to provide basic access to transcribed speech corpora.

For storing the internal backup copies of each resource, we use IDA⁶, a research data storage service organized by the Ministry of Education and Culture in Finland. IDA is offered free of charge to Finnish universities, universities of applied sciences and state research institutes. The service allows researchers and teams not only to save, organize and share their research datasets but also to freeze the data, i.e., to describe datasets and to store them in an immutable state for long-term archiving.

2.2 Access Rights

The Language Bank aims to provide resources as openly as possible. Many resources can be made publicly available (CLARIN PUB license category). However, access restrictions may be necessary in case the resource includes, e.g., copyrighted content or personal data that should be protected.

Some resources are licensed for academic use only (ACA), and they may be accessed by signing in with credentials issued by the user's home institution. Furthermore, the Language Bank is able to distribute resources under restricted licenses (RES), in which case users can apply for individual access rights in the Language Bank Rights (LBR) service⁷. LBR currently supports federated login and user identities via CLARIN⁸ or Eduuni⁹. The Language Bank uses the common CLARIN licensing framework, with some local adjustments¹⁰. Unless the original material has been previously available under a public license, the licenses of individual resources in the Language Bank are based on agreements with the rightholders and, in the case of resources that contain personal data, with the data controllers.

In some cases, it is possible to offer several variants of the same resource under different licenses. For instance, since speakers might be identifiable based on their voice, audio speech recordings often need to be protected, e.g., by restricting access to them. However, it may be possible to make the anonymized or pseudonymized transcripts available under a less restricted license for specific purposes where access to the audio is not needed.

The Language Bank Rights system is based on REMS (Resource Entitlement Management System)¹¹, an open-source electronic tool developed by CSC for the management of access rights to research data. Researchers can log into the system by using the user credentials provided by their home organisation. The researcher selects a resource for which access rights are applied, fills in an electronic application form, and agrees to the terms of use for the dataset in question. The application can then be circulated via LBR to the rightholder's representative for approval. LBR can also be used, e.g., for contacting the supervisor of a student applicant in case their endorsement is required before access can be granted to a specific resource. From LBR, it is also possible to obtain reports on applications and approved access rights.

2.3 Documentation

While developing our publishing processes, we have paid attention to the systematic documentation of the resources as well as to improving their findability. Corpora can be found either on the list of published resources or on the list of forthcoming resources in the portal.

Over time, several versions and variants have been published of individual resources on different platforms. To help the users find out which version is the most relevant one for them, we use the so-called resource group pages for documenting all the versions and variants of a given resource as a group. The resource group pages may include additional resource-specific instructions that cannot be included

⁶ <https://ida.fairdata.fi>

⁷ <https://lbr.csc.fi>

⁸ <https://www.clarin.eu/content/clarin-identity-provider>

⁹ <https://info.eduuni.fi/en/services/eduuni-id>

¹⁰ <https://www.kielipankki.fi/support/clarin-eula/>

¹¹ <https://www.csc.fi/remms-kayttovaltuuksien-hallintajarjestelma>

in the individual metadata records (especially when there are many versions). The resource group page of a given resource lists all available versions of the resource, including links to their metadata records, their access locations, and further information. A link to the resource group page can be found in the metadata record of each version of the resource.

Following the example of CLARIN Resource Families¹², we also offer a portal page where the resource groups in the Language Bank are categorized under CLARIN-style families.

3 Challenges and Goals

While implementing and developing our publishing pipeline, we aim to meet the needs of the users as well as to improve our internal workflows. Resources should be offered in consistent and interoperable formats and they should be easy to find and to process by researchers and research groups.

More than 250 resources are currently available via the Language Bank of Finland. About 100 resources are listed as forthcoming, and more are added every month. Before implementing the publishing pipeline, each team member involved in the process of publishing resources had their own workflows and scripts for converting data. This often resulted in slight inconsistencies in the published resources. In addition, it was not easy to monitor the state of each resource within the publication process. Certain tasks, such as parsing the data for publication in Korp, were carried out by only one person in the team, making processes very dependent on this person's availability and time.

The process of publishing an individual corpus usually involves 3–4 people in the Language Bank. In case of an exceptionally simple and well-described dataset with no licensing issues, it does not take more than one or two working days to publish the source data for download. If intense license discussions and several different means of publication are required, the process can take up to 60 working days.

Our aim is to perform faster. Ideally, the process should enable us to publish resources within a shorter time frame and to reduce the amount of time that a certain resource ends up waiting in the pipeline for the next processing steps. To make the workflow more efficient, it is important to be able to monitor the status of a resource during the publishing process and to share tasks and knowledge within the team. For this purpose, we have collected a list of the specific tasks that are addressed during the publication process of most resources.

4 Tasks within the Publishing Pipeline

For each new resource, we maintain a checklist¹³ of the tasks in the shared pipeline that are relevant for the resource in question. The list is used for keeping track of the status of the resource during the publishing process. Some tasks on the list are mandatory for all types of resources, whereas others are applicable to specific types only. According to the type of the task, which can be for example administrative or technical, work can be assigned to a person with the required skills.

The tasks can be classified into different groups, as shown in figure 2, starting at entering the new resource and clearing the license, and finishing at publishing the resource in the different channels and formats. Grouping the tasks thematically as well as chronologically helps us keep track of the status of each resource and to make the checklist of tasks clearly laid out. It is to be noted, however, that the order of the events in the publication pipeline is not a strict timeline but rather a dependency structure. For instance, data acquisition and processing may in some cases start before the deposition license agreement with the data provider is signed, whereas the corpus cannot be published unless the license is cleared.

A ticketing system helps in managing and monitoring the publication processes of individual resources. For this purpose, we use Atlassian JIRA. Currently we are looking for a more convenient way

¹² <https://www.clarin.eu/resource-families>

¹³ <http://urn.fi/urn:nbn:fi:lb-2023032703>

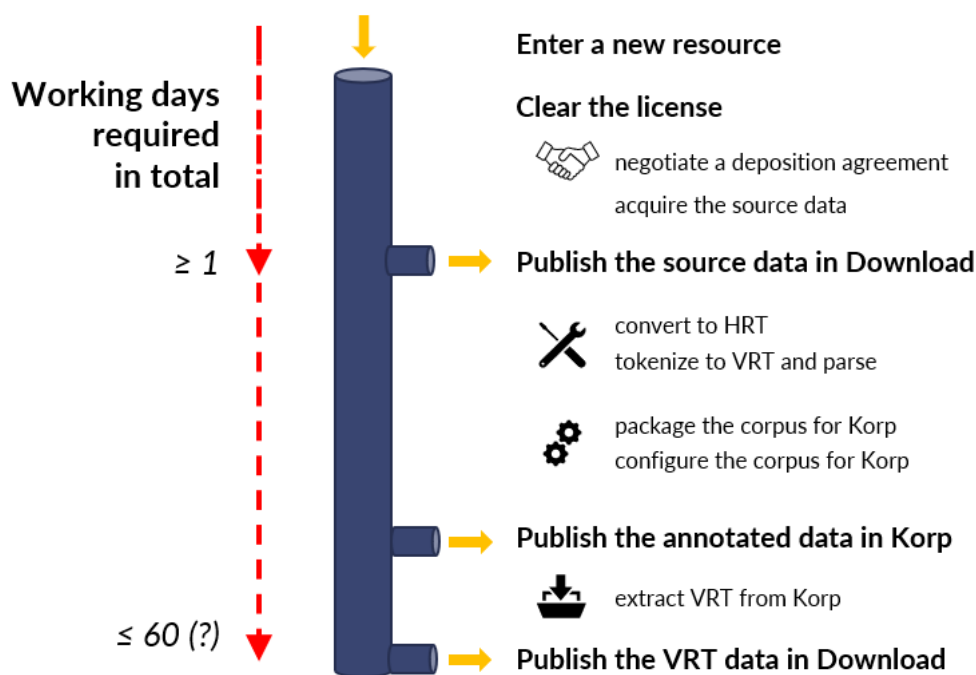


Figure 2. Tasks within the publishing pipeline

of implementing the list of tasks to the pipeline, so that tasks could be automatically created for each incoming resource and made more accessible to the team.

4.1 Entering a New Resource to the Pipeline of the Language Bank of Finland

When a researcher or a research group creates a new resource that they wish to make available to other researchers or publicly, they are first asked to submit the most important details regarding the resource by filling in an e-form¹⁴. The Language Bank then creates a preliminary metadata record on the local META-SHARE repository¹⁵, where the metadata of all resources available via the Language Bank are currently maintained. The preliminary metadata are checked together with the depositor. The details can be updated and amended later. In order to make sure that the metadata records meet our quality standards and remain consistent, the editing rights to the metadata records are restricted to a few people with the required expertise.

The metadata records in the META-SHARE node of the Language Bank are harvested by our OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) metadata provider that transforms the native META-SHARE records to CLARIN's CMDI format. The metadata are then available to external services, such as the Virtual Language Observatory¹⁶ maintained by CLARIN. Our local provider validates the records before and after the transformation, using the corresponding XML schemas, to avoid providing malformed metadata and to inform our metadata maintainers about faulty records. The XSLT script transforming the data to the CMDI format was provided by the implementors of the META-SHARE format.

¹⁴ <http://urn.fi/urn:nbn:fi:lb-2021121422>

¹⁵ <https://metashare.csc.fi>

¹⁶ <https://vlo.clarin.eu>

Reference instructions: AVOID

Please cite the language resource as follows:

Kinnunen, T., Hautamäki, R. G., Sahidullah, M., Hautamäki, V., Werner, S., & Bentz, M.. *Corpus of Age-related Voice Disguise (AVOID)* [speech corpus]. Kielipankki. Retrieved from <http://urn.fi/urn:nbn:fi:lb-2018060621>

Show: [\[Bibtex\]](#) [\[Zotero\]](#)

[Search for references to the language resource in Google Scholar](#)

Figure 3. An example of reference instructions

For publications, researchers may need a persistent reference to their resource before it is made available by the Language Bank. Since unofficial links should be avoided in citations, the Language Bank assigns a persistent identifier (PID) to the metadata record as soon as the resource exists and has been sufficiently well described. Each version of the resource gets a metadata record and a PID of its own. For instance, the downloadable source version will have a different PID from the Korp version of the resource. The metadata PID is the citable and primary identifier of the resource version. For internal use, we also assign PIDs to the access location, to the license pages and to the resource group pages. The Language Bank uses the URN system for generating PIDs. For details on how the Language Bank uses PIDs, see Matthiesen and Dieckmann, 2019.

At this point, the resource is added to the list of forthcoming resources on the website of the Language Bank. The reference instructions will then be automatically generated and displayed on demand. A link to the reference instruction is included in the META-SHARE record of the resource and in the resource group page. The reference instruction (for an example, see figure 3) currently provides the user of a resource with the names of the author(s) listed in the order required by the rightholder in the deposition agreement, the publication year of the dataset, the full title of the corpus, the type of the dataset (e.g., [text corpus]), the name of the centre or repository that is maintaining the resource for access (The Language Bank), and, lastly, the citable PID of the resource in question.

In addition to the full title, each version of a resource is also labelled with a short name that is included in the corresponding metadata record. The short name typically consists of a stem, derived from the official name or acronym of the corpus, an optional version number, and a suffix that carries information about the type of the resource variant (cf. figure 4). The short names can be used as keys for keeping the resource versions distinct from each other within various services, e.g., in the names of portal pages and in the file and folder names of downloadable corpora. The short names are also used as keys for generating the citation instructions in the web portal. As shown in the examples in figure 4, the members

Abbreviation	Name and metadata
wanca2016-korp	Wanca 2016, Korp Version
wanca2016-src	Wanca 2016, source
wanca2016-vrt	Wanca 2016, VRT

Figure 4. Examples of the short names of three resource variants within the resource group 'wanca'

of the Wanca resource group can be conveniently kept together since the stem 'wanca' is systematically included in the short names of the different versions and variants of the resource.

4.2 Clearing the License for the Resource and Acquiring the Source Data

Unless the resource has previously been published under an open license, the Language Bank and the depositor negotiate on the license for distributing the resource. In case there are no legal reasons for restricting the use of the resource to specific purposes, users or user groups, it is preferable to select an open license. However, language resources may contain for instance copyrighted text, and it is necessary to ensure that the depositor has sufficient rights to share the material via the Language Bank. Moreover, the resource (or at least certain parts of it) may include personal data, which must be considered so that the appropriate safeguards can be applied when the material is stored and processed by the Language Bank and by the end-users in their own research projects.

In case the resource contains copyrighted content, additional steps may be needed to obtain permissions from the copyright holders. These permissions are usually requested by the researcher who wishes to deposit the resource, but the Language Bank can offer support for formulating the requests.

If the resource includes personal data, the data controller, responsible for the original purpose of processing the data, is involved in the deposition agreement. In this case, the end-user license will include the condition +PRIV, and all users who access the resource via the Language Bank will be required to comply with the resource-specific data protection terms and conditions.

The Language Bank uses a generic deposition license agreement template¹⁷. In order to discuss the details, a meeting with the depositor is often needed. When an agreement is reached, the end-user license is published in the portal. Using PIDs, the metadata record will refer to the license page and vice versa.

After receiving the source data from the resource depositor, the data are checked for format and validity, and a description of the contents is added for internal use. A backup copy of the data is stored in IDA.

4.3 Publishing the Source Data in Download

Since the data conversion process tends to take time, the first version of a corpus to be published in the Language Bank is usually the source data that is made available for download. In this version, the original content is not modified. However, the metadata and license information must be available and up to date.

A PID is added to the metadata record, and a resource group page, which also gets a PID, is created and linked with the corresponding metadata. The source version of the resource, and possibly some other foreseen versions of the resource, are added to the list of forthcoming resources in the portal, to keep the corpus owners and the potentially interested researchers informed. The attribution details are added to the metadata record. Furthermore, PIDs are assigned to the license page and the download location.

¹⁷ <https://www.kielipankki.fi/support/dela/>

To prepare the resource for download, the source data is packaged into one or more zip files as agreed with the corpus depositor, including a README text file that contains basic information on the resource and a LICENSE text file offering information on the access rights for this resource. In case the license of the resource is in the RES category, a record is created on the LBR system in order to be able to control access to the download location. Similarly, if the license is ACA, academic user login will be required to download the resource. After a successful check of the quality and accessibility of the uploaded zip packages, the metadata record and the resource group page can be updated with the access location PID.

To finalize the publishing of a resource, it is moved from the list of forthcoming resources to the list of published resources in the portal. A news item is published in the portal to inform interested researchers about the new resource. The depositor is informed about the publication as well. The download package is uploaded to be stored in IDA, and in selected cases, for example when wide use of a large resource is anticipated, the unpacked source data is also made available in CSC's computing environment.

4.4 Publishing the Data in Korp

Our goal is to make data accessible to the user in a uniform format, converted from various source formats. For this purpose, we use VRT, which is the input format for the IMS Open Corpus Workbench (CWB) software (Evert and Hardie, 2011) underlying Korp. The data is first tokenized, and annotations are inserted in order to include any descriptive information available in the source data, such as the dates, locations and authors of the individual texts. The data is then lemmatized, tagged with parts of speech and/or parsed, depending on the automatic annotation tools available for the language in question. The data can also be extended with additional annotations, such as name annotations, sentiment annotations and identified languages.

The first steps of publishing the Korp version of the resource are similar to those of the downloadable versions. A metadata record for the Korp version is created or updated and PIDs are assigned to the metadata record and to the access location. In case the license of the resource is RES, an LBR record is created.

The format of the original data tends to vary between corpora. It can be for example plain text, PDF, RTF, or an XML format such as TEI. For Korp, PDF documents are first converted into text files. Unless the source data is already tokenized, the first aim is to convert this data to a simple, XML-style format which must be UTF-8-encoded Unicode. An example of this format, which we call HRT (*HoRizontal Text*), is shown in figure 5. The conversion is carried out with tailor-made scripts and it can often be the most time-consuming step, depending on the format of the original data. The basic idea is to segment the content of the original files so that the plain text is inside text and paragraph tags, which can include descriptive attributes. These files with a relatively simple structure are then used as input for further processing tools. The next step is the tokenizing process where the paragraphs are segmented into sentence elements and tokens. The output format of the tokenizer is VRT, which we have extended with a comment that provides names for the otherwise positional attributes of tokens. An example of the VRT format can be seen in figure 6.

It is possible for the Language Bank to apply further tools on the VRT data to add any desired annotations, while preserving the sentence and token boundaries and previous annotations. For instance, information about the languages used in the text can be added by running a language identifier such as HeLI-OTS (Jauhiainen and Jauhiainen, 2022) that includes language models for 200 languages.

For Finnish and other languages with a parser and named-entity recognizer available, the parsing process is carried out on the validated VRT data. For years, we have been using an early version of the Turku dependency parser for Finnish, developed by the Turku NLP group and adapted for VRT. We are currently adopting their new neural parser¹⁸ along with the Universal Dependencies annotation model.

A single script calling several other scripts handles the processing of VRT files to create a Korp corpus package containing the CWB data files and the Korp MySQL database import files. The resulting package is then installed on the Korp server, and a Korp corpus configuration is added with the information on the corpus and its annotations (attributes). The corpus configuration determines where and how the corpus is shown in Korp. The configuration is first added to a test instance of Korp. When

¹⁸ <http://turkunlp.org/Turku-neural-parser-pipeline/>

```

<text binding_id="1377028" date="1986" datefrom="19860101" dateto="19861231">
<paragraph id="0">
mahdollista.
</paragraph>
<paragraph id="1">
Malminkartano on kuuluisa linnastaan, jota eräässä
kulttuurihistoriallisessa asiantuntijalausunnossa on kehuttu Suomen
komeimmaksi kartanorakennukseksi. Se valmistui 1885 entisen puutalon
tilalle, jonka alakerros oli peräisin 1600-luvulta.
</paragraph>
<paragraph id="2">
Arkkitehtina oli F.A. Sjöström ja tyyli on Suomen maaseudulla melko
harvinaista uusrenessanssia, jonka arvostus aleni pian siinä määrin,
että rakentajan oma tyttärentytär puhui "maun
rappiokaudesta". Myöhemmin on ymmärtämys sitä kohtaan kasvanut.
</paragraph>
</text>

```

Figure 5. An example of the HRT format

the test instance meets the expectations, the configuration is copied to the production Korp, where the corpus is published as a beta version.

The new corpus is announced in the Korp news desk as well as in the portal. The beta status is removed after two weeks unless requests for changes are received during this period. Finally, a copy of the Korp corpus package is stored in IDA.

Although this approach is time-consuming, it has been designed so as to ensure the consistency and interoperability of the published resources. It is important to preserve as much of the information in the original data as possible, be it structural information or metadata at various structural levels of the data. We are able to reach this goal by using tailor-made scripts for converting the source data to HRT.

All the generic and tailor-made scripts used for processing corpora are published openly on GitHub¹⁹.

4.5 Publishing the VRT data in Download

After publishing a resource in Korp, the VRT data is usually extracted from Korp and published in the download service, in order to provide consistent versions of the data via both channels. The VRT version of a resource is published in the download service in the same way as the source data. For the VRT version, a separate metadata record is created, and the corpus version is added to the resource group page.

We decided to make the VRT versions of the data available for the users, since we believe that VRT can be useful for further processing. The VRT format²⁰ is simple and human readable and easy to process and to transform. It is a combination of one-word-per-line (vertical) format and simple XML markup. In the future, we also intend to offer tools for easy conversion from VRT to other formats.

4.6 Testing and quality control

When a corpus is ready to be published in a given means of publication, the final step before the actual publishing is quality control, i.e., a testing procedure is required. Ideally, the testing should be carried out by a member of the team not involved in the processing of the resource in question. Testing procedures are tailored for download and Korp separately. They also differ between (versions of) resources with different access rights. The accessibility on the different platforms is tested, and it has to be made sure that ACA and RES restrictions work as expected.

Our testing procedures are still under development. Our aim is to have a catalogue of test cases available, covering what should be tested from the user's perspective as well as taking internal needs like archiving and documentation into account. A comprehensive checklist covering all the various cases together with clear guidance and possibly screenshots should enable every member of the team to take

¹⁹ <http://urn.fi/urn:nbn:fi:lb-2023032701>

²⁰ <http://urn.fi/urn:nbn:fi:lb-2023020121>


```

<!-- #vrt positional-attributes: ref word lemma pos msd deprel dephead -->
<text binding_id="1377028" date="1986" datefrom="19860101" dateto="19861231">
<paragraph id="0" sum_lang="|xxx:1|">
<sentence id="0" lang="xxx" lang_conf="3.7488587">
1   mahdollista   mahdollinen   A       NUM_Sg|CASE_Par|CMP_Pos ROOT    0
2   .             .           Punct   _       punct    1
</sentence>
</paragraph>
<paragraph id="1" sum_lang="|fin:7|xxx:1|">
<sentence id="1" lang="fin" lang_conf="1.3936844">
1   Malminkartano malmi|kartano N       NUM_Sg|CASE_Nom|CASECHANGE_Up nsubj-cop 3
2   on            olla        V       PRS_Sg3|VOICE_Act|TENSE_Prs|MOOD_Ind cop        3
3   kuuluisa     kuuluisa   A       NUM_Sg|CASE_Nom|CMP_Pos ROOT    0
4   linnastaan  linna     N       NUM_Sg|CASE_Ela|POSS_Px3 nommod    3
5   ,            ,          Punct   _       punct    11
6   jota        joka      Pron   SUBCAT_Rel|NUM_Sg|CASE_Par rel        11
7   eräässä    eräs      Pron   NUM_Sg|CASE_Ine det        9
8   kulttuurihistoriallisessa kulttuuri|historiallinen A       NUM_Sg|CASE_Ine|CMP_Pos amod    9
9   asiantuntijalausunnossa asian|tuntija|lausunto N       NUM_Sg|CASE_Ine nommod    11
10  on            olla        V       PRS_Sg3|VOICE_Act|TENSE_Prs|MOOD_Ind auxpass   11
11  keuhattu    kehua     V       NUM_Sg|CASE_Nom|VOICE_Pass|PCP_PrfrPrnc|CMP_Pos rcmmod    4
12  Suomen      Suomi    N       SUBCAT_Prop|NUM_Sg|CASE_Gen|CASECHANGE_Up poss      13
13  komeimmaksi komea     A       NUM_Sg|CASE_Tra|CMP_Super1 amod      14
14  kartanorakennukseksi kartano|rakennus N       NUM_Sg|CASE_Tra nommod    11
15  .            .          Punct   _       punct    3
</sentence>
</paragraph>

```

Figure 6. An example of the VRT format

care of the quality control. Also validation scripts can help, where feasible, as they produce meaningful reports and are usually easy to run.

5 General Discussion

While using and developing further our publishing pipeline for a few years now, we have discovered advantages and disadvantages in our approach. Especially the time and effort required from our side to publish a resource can be seen as problematic in this respect.

After introducing the resource group pages, the resource families page, license pages for public resources, pages for the resource-specific data protection terms and conditions, etc., the number of administrative tasks has increased to some extent. The actual working time spent on publishing one version of a resource might not have changed significantly. However, the waiting time within the pipeline has diminished, since all team members can at least in principle perform the required tasks, due to the shared scripts and documentation. It is now much easier for us to manage the tasks and to monitor the publishing process for each resource. The quality of the published packages and the findability of the corpora has also improved.

We aim to offer resources in uniform formats, to preserve the information in the original data to a maximal extent, to make resources available while respecting the potential restrictions regarding access rights, and to ensure the quality, findability and accessibility of the published resources. These goals may justify the current active role of the Language Bank in the publication workflow. Nevertheless, by developing our processes and by making them more automatic, it will be possible for the resource creators to participate more and more actively in the publication of their data.

The main focus of the publishing pipeline is currently on processing text corpora via Korp and on creating download packages. Ever since the former LAT system was taken out of use in the Language Bank at the end of year 2020, we have been looking for a suitable replacement service that would enable users to query, to browse and to access speech and sign language corpora that may include audio and/or video files as well as time-aligned annotations in multiple annotation tiers. Currently, most of the speech and sign language corpora available in the Language Bank are offered via the download service only, and the users are instructed to use locally installable GUI tools for querying the annotations and for analysing the multimedia content, e.g., ELAN²¹ (MPI in Nijmegen) for audio and video, or Praat²²

²¹ <https://archive.mpi.nl/tla/elan>

²² <http://www.praat.org/>

(Boersma and Weenink, 2023) for audio and digital signal processing. It is hoped that new solutions can be developed so as to provide a higher level of service to audio and video corpora in the future.

The Language Bank is currently working towards building a centralized database that includes all the public as well as internal metadata of each resource, their status in the pipeline and the license details. We intend to use the database for generating and maintaining the public metadata records, the README documents included in the download packages, the license pages, the resource group pages, the listings of currently available and forthcoming corpora, etc. The database could also be plugged in with the PID generator and be used to automatically provide the citation instructions. By storing and maintaining all the resource information in one place, we would be able to reduce the need for manually copying and pasting data between a number of documents and web pages. Moreover, the users could be allowed to submit preliminary information directly to the database. The resource database will help us minimize human errors and administrative delays in the resource publication process.

We believe that the different CLARIN centres should be able to collaborate with each other in developing their technical as well as administrative practices. In order to gain a better understanding of what sort of expertise is already available within CLARIN, an overview of the main services, platforms and means of user support that are offered by each CLARIN centre would be useful. We are planning to conduct a survey in collaboration with the Standing Committee for CLARIN Technical Centres and the CLARIN User Involvement Committee.

6 Conclusions

Currently, the Language Bank of Finland provides researchers with access to over 250 resources, and many more are forthcoming. The licensing and publishing process of each resource takes time and effort and requires various kinds of expertise. Based on our experience, we have identified a number of tasks that are relevant when publishing most types of resources, resulting in a checklist and modular documentation²³ offering instructions for the individual tasks. Although this pipeline is still under development, the general workflow has already proven useful for managing and monitoring the publication process more efficiently.

We aim to automate and document our processes even further to enable resource depositors to take a more active role in preparing their data. The pipeline should also be extended in order to make it more convenient for users to discover and to share tools via the Language Bank. We believe that by comparing and sharing good practices with other CLARIN centres, it is possible to support researchers even better.

References

- Paul Boersma and David Weenink. (2023). Praat: doing phonetics by computer [Computer program]. Version 6.3.06, retrieved 31 January 2023 from <http://www.praat.org/>
- Lars Borin, Markus Forsberg and Johan Roxendal. (2012). Korp – the corpus infrastructure of Språkbanken. *Proceedings of LREC 2012*. Istanbul: ELRA, pages 474–478.
- ELAN (Version 6.4) [Computer software]. (2022). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from <https://archive.mpi.nl/tla/elan>
- Stefan Evert and Andrew Hardie (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. *Proceedings of the Corpus Linguistics 2011 conference*. University of Birmingham, UK.
- Tommi Jauhiainen and Heidi Jauhiainen. (2022). HeLI-OTS 1.3 (1.3). Zenodo. <https://doi.org/10.5281/zenodo.6077089>
- Martin Matthiesen and Ute Dieckmann. (2019). A PID is a Promise – Versioning with Persistent Identifiers. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 103–112.

²³ <http://urn.fi/urn:nbn:fi:lb-2023032702>

TEI and Git in ParlaMint: Collaborative Development of Language Resources

Tomaž Erjavec

Dept. of Knowledge Technologies
Jožef Stefan Institute
Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Matyáš Kopp

Faculty of Mathematics and Physics
Charles University
Prague, Czech Republic
kopp@ufal.mff.cuni.cz

Katja Meden

Dept. of Knowledge Technologies, Jožef Stefan Institute,
Jožef Stefan International Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia
katja.meden@ijs.si

Abstract

This paper discusses the encoding, validation and development of language resources in the completed ParlaMint I and on-going ParlaMint II CLARIN projects, which centre on the collaborative development of a large set of interoperable corpora of parliamentary proceedings. It focuses on the encoding of ParlaMint corpora, the GitHub development platform, and the evaluation of their use by project partners. We introduce the use of TEI for the encoding guidelines and validation schemas. We motivate and explain the use of Git and GitHub to develop and maintain the encoding schemas, validation and conversion scripts and samples of the corpora. The paper also presents the results of a survey on the use of TEI and Git in the ParlaMint projects among the project participants. Overall, participants were mostly positive about their experience with TEI and Git, although some difficulties were reported. These will serve as a basis for further TEI and Git optimisation in ParlaMint.

1 Introduction

ParlaMint is a CLARIN ERIC supported project¹ which, among other tasks, aims to produce a set of comparable and richly annotated corpora, involving a joint effort of a large number of partners. The concluded ParlaMint I project (2020–2021) already developed corpora containing transcriptions of the sessions of 17 European national parliaments in the time-span 2015–2021 (Erjavec et al., 2022). These corpora are about half a billion words in size, contain rich metadata on 11 thousand speakers, and are linguistically annotated. The on-going ParlaMint II project (2022–2023) plans to extend existing corpora with newer data and add corpora for 14 new, also regional, European parliaments. It will also enhance the corpora by providing machine translations to English, and, for a selected subset of corpora, add speech data, as well as work on the wider use of the corpora.

With a largely bottom-up project involving many partners, parliamentary systems and different sources but aiming to produce a large set of highly comparable corpora, it is important to have robust encoding guidelines, automated validation and a scalable and flexible data workflow. ParlaMint I was already using TEI and Git to achieve these goals, and the first tasks in ParlaMint II were to reevaluate and extend these aspects of the project.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

¹<https://www.clarin.eu/parlamint>

The paper discusses the project’s use of TEI, including developing the encoding guidelines, the formal XML schema, and the validation and conversion scripts (Section 2); the use of Git and GitHub as an open and controlled development environment (Section 3); the results and analysis of a survey conducted among the partners in ParlaMint II on their familiarity, use, and suggestions as to the use of TEI and Git (Section 4), ending with a summary and directions for further work (Section 5).

2 Encoding the ParlaMint Corpora

The definition of a common encoding for parliamentary corpora has a long history in the context of CLARIN, starting with the CLARIN Travelling Campus “Talk of Europe” where, in 2014 and 2015, three hackathons were organised using European Parliament proceedings curated as linked open data. In addition, interdisciplinary workshops on working with parliamentary records and co-located with the LREC conference were organised (2018, 2020, 2022) under the guidance of CLARIN. Finally, the Parla-CLARIN recommendations for encoding parliamentary corpora (Erjavec and Pančur, 2021)² were proposed at the “CLARIN ParlaFormat Workshop” in 2019.

Parla-CLARIN is a customisation of the TEI Guidelines (TEI Consortium, 2022)³. A TEI customisation is specified in a TEI ODD (One Document Does it All)⁴ document, which serves a double function: it contains the prose guidelines, as well as the formal schema of the customisation, using the TEI ODD schema specification language. With the TEI XSLT stylesheets the prose guidelines can be converted to HTML for reading, while the ODD schema specification is converted into one of the standard XML schema languages, such as the ISO standard RelaxNG, and such an XML schema is then used for formal validation of the corpora. The design of the Parla-CLARIN recommendation was inspired by previous similar efforts, in particular the TEI Lex-0 encoding recommendations for dictionaries (Tasovac et al., 2018)⁵ and the TEI schema for the multilingual ELTeC corpus containing 100 historical novels for a number of languages (Burnard et al., 2021; Schöch et al., 2021)⁶.

Although the ParlaMint corpora conform to the Parla-CLARIN schema, we required, in order to ensure interoperability, a much more constrained encoding than the quite general one of Parla-CLARIN. To this end, we have, in ParlaMint I, developed a bespoke RelaxNG schema without using the ODD mechanism. The advantage of this approach is that the schema expresses exactly the kinds of constraints that we wished to make, while the disadvantage is that there were no guidelines accompanying the schema, which is, formally, not even TEI, exactly because the schema was not derived from a TEI ODD. For these reasons, we developed, in ParlaMint II, a ParlaMint ODD, which contains the prose guidelines and as well as the formal schema. The ODD schema allows only the elements and attributes that we wish to have in ParlaMint, however, still with richer content models than those required, as it is quite difficult (short of completely re-specifying the content models of all the elements, or introducing a special namespace) to forbid the multitude of element nestings otherwise allowed by TEI. We have also worked on the documentation of individual elements and attributes in the ODD schema, i.e. changing the default glosses and examples of use of the elements as they appear in the TEI to ParlaMint specific ones.

The ParlaMint schema (either the bespoke or the ODD-derived RelaxNG) is only the first step in the validation of the ParlaMint corpora. We have also developed an XSLT script that performs validation regarding the textual content of some elements and checks that redundant metadata (i.e. metadata which is encoded in several places in the corpora but which makes it easier to inspect or process the corpora further) is not contradictory. Another script checks that all cross-references are resolvable, i.e. that a corpus does not contain broken internal links.

²<https://clarin-eric.github.io/parla-clarin/>

³<https://tei-c.org/guidelines/P5/>

⁴<https://tei-c.org/guidelines/customization/getting-started-with-p5-odds/>

⁵<https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>

⁶<https://github.com/COST-ELTeC>

Furthermore, as the corpora are converted to other formats, such conversions can also expose various errors in the corpus. One of the down-stream formats is CoNLL-U, and the Universal Dependencies validation tool⁷ is used to check the validity of the linguistic analyses. The corpora converted to vertical files are mounted on CLARIN.SI concordancers, where the conversion from TEI, the corpus compilation log, and an analysis of corpora on the concordancers can also reveal problems.

The mark-up of a ParlaMint corpus is rather complex compared to most linguistically annotated corpora, first because of extensive metadata about the speakers and the political organisations (e.g., lower/upper house, political party), with temporal attributes attached to the metadata. We have also tried to retain many aspects of the original transcripts, in particular transcribers' comments, which are encoded in several ways, depending on their content. However, this complex encoding is explained and exemplified in the ParlaMint encoding guides and samples from existing corpora are directly viewable via GitHub, all with the intention of making it easier to come to grips with the encoding system,

3 Using Git and GitHub

Git has become the revision control system of choice for many software development projects, and has also proved its worth in the (collaborative) development of language resources, with the most prominent example being the Universal Dependencies treebanks and annotation guidelines (de Marneffe et al., 2021)⁸. It has also been used for the development of TEI customisations, e.g. the already mentioned TEI Lex-0 and ELTeC, for the latter used not only of the schema, but of the corpora as well.

Apart from support for collaborative development with transparent versioning and attribution, and simple comparisons of files, Git hosting platforms, such as GitHub, support social media aspects of development, in particular posting and discussing issues, commenting on commits or pull requests, and a Wiki space. It is also possible to directly publish the documentation of a project using GitHub Pages. Finally, running scripts at a particular point in the Git workflow is supported by GitHub Actions. All these features lead to a more controlled and better documented development process.

We had already used GitHub for the development and publishing of the Parla-CLARIN schema, where the TEI ODD is maintained on GitHub⁹, the guidelines are published as GitHub pages¹⁰, and technical instructions for using or further developing the schema are available on the GitHub Wiki¹¹. In ParlaMint I, as well, the project development was to a large extent done on GitHub¹². The Git repository contained the latest RelaxNG schemas for the corpora and the complete validation or transformations scripts, written mostly in XSLT (and some Perl). Problems with the proposed encoding schema were often discussed through GitHub issues, while problems with individual corpora were communicated mostly by email.

Git(Hub) is, of course, not the only revision control platform available, so the question is, why we have chosen to use exactly this option. The reasons are, to a large extent, pragmatic in nature. We are familiar with GitHub, so it is easier to use the platform we know; in ParlaMint I we already used various GitHub's features, so it also made sense to use the same platform in ParlaMint II, because the ParlaMint I partners were already familiar with the platform and established workflow there; finally, GitHub is owned by a very large company, so we can be reasonably sure it will be maintained in the foreseeable future.

But while Git(Hub) is well suited for developing, storing and publishing software tools, schemas, and even quite large hand-annotated corpora, the complete ParlaMint corpora are, in

⁷<https://github.com/UniversalDependencies/tools>

⁸<https://universaldependencies.org/>

⁹<https://github.com/clarin-eric/parla-clarin/>

¹⁰<https://clarin-eric.github.io/parla-clarin/>

¹¹<https://github.com/clarin-eric/parla-clarin/wiki>

¹²<https://github.com/clarin-eric/ParlaMint>

practice, somewhat large to be stored in Git, and, even more so, GitHub. Apart from the sheer size of the corpora (over 240 GB for ParlaMint I) and large number of files (almost 160,000), this is also due to the fact that, say, a new round of automatic annotation changes almost all files, making such a commit a very slow process: an experiment with complete ParlaMint I corpora showed the initial staging and commit to take approximately 4 hours.

It should also be noted that we also use GitHub’s social network features, so it is very helpful to be able to view the content of files as rendered in a web browser directly from GitHub. While we haven’t found the exact file size limit in GitHub’s documentation, our experience shows that it is possible to only view rendered files that are smaller than 2 MB. Another reason for working with a smaller portion of corpora is that annotating one corpus can take a few days, and the errors usually appear in most files so a small sample is sufficient for debugging and discussing doubts in encoding. Therefore we here opted for a compromise, namely, we developed a script that extracts only small samples from individual corpora, and maintain only these samples, also in derived formats, on GitHub. The script reduces the number of files as well as the amount of XML content in particular files so these samples can be directly viewed on GitHub. This gives an impression of how the corpora are structured and makes development more manageable because it is possible to refer to particular parts of files in issues.

The complete set of corpora is then made available only for a major release and deposited in the CLARIN.SI repository.

In ParlaMint II, the first step was to update the Parla-CLARIN GitHub to reflect the ParlaMint best practice, while the ParlaMint GitHub was extended with pages¹³ which are used to publish the ParlaMint encoding guidelines. Apart from submitting the complete corpora, all the communication was done through GitHub issues, rather than via email, so that problems are documented, can be discussed and the solution linked to a commit.

At the time of writing, over 300 issues, with over 1,700 posts have been opened, with the majority already resolved. 52 different GitHub users contributed to creating an issue, pull request or responding to either. The most discussion was, unsurprisingly, on issues related to problems when merging the pull requests submitting samples for new corpora, with one having almost 70 comments and replies.

As we cooperate with many partners that are supposed to add their sample data with pull requests, a validation procedure for newly inserted data using GitHub Actions¹⁴ has also been developed. Furthermore, when a pull request with valid data is merged into the correct branch, the TEI files are sampled, and derived formats are added to the repository. This approach has several benefits: there is no need for the partners to carefully sample the data themselves, they do not need to compile the derived formats, and we can be sure that the derived files are always up to date with corresponding TEI files.

For local validation and conversion of the complete corpora either by a partner or centrally, we have developed a validation procedure that uses the Unix `make`¹⁵ tool. The Makefile is self-documenting for easier use, i.e. running `make` without arguments prints a list of the available targets, which e.g. check installed prerequisites, validate the corpus against the ParlaMint and Parla-CLARIN schemas, perform advanced content validation, and convert a corpus to derived formats. Finally, the Perl wrapper program is used to prepare a ParlaMint corpus for distribution: it finalizes the corpus header, runs all the validation steps, converts the corpus to derived encodings and packs the corpus, both as a “plain text” corpus and as a linguistically annotated one.

¹³<https://clarin-eric.github.io/ParlaMint/>

¹⁴<https://docs.github.com/en/actions>

¹⁵<https://www.gnu.org/software/make/>

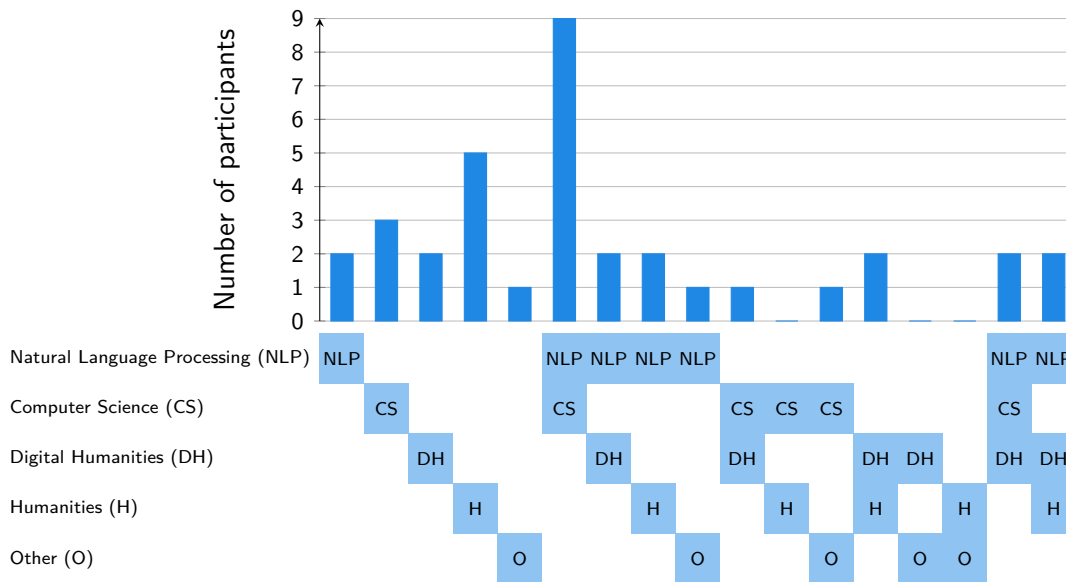


Figure 1: Distribution of research background combinations by individual participants (multiple answer question, 35 participants, 61 answers in total)

4 The Survey on TEI and Git

Since the development of the ParlaMint corpora involved numerous partners with varying degrees of familiarity with TEI and Git, we decided to solicit their feedback to assess the development pipeline and identify opportunities to further improve the workflow. To this end, we designed a questionnaire that included three main sections: an introductory section to identify the key characteristics of our project partners, a section on TEI encoding with questions about the encoding process and experiences with TEI, and finally a section on Git and GitHub that included questions about their familiarity with Git and GitHub. The survey ran from January 3 to January 20, 2023.

The analysis of the survey results is presented separately for TEI and Git. In both cases, we first present the results of the survey, based on which we conducted further analysis to explore various relationships between participants (and their research backgrounds) and their experiences with TEI before and after project work on ParlaMint.

Most responses were submitted by DK, SI (each 4 completed surveys) and IT (3 surveys), while AT, BA, GR, HR, HU, PL, RS, UA shared 2 completed surveys per country. We also received responses (1 completed survey per country) from BE, BG, EE, ES-PV¹⁶, FR, IS, LV, NL, NO, PT, RO, SE, TR, giving us a total of 35 responses from 24 countries of 31 ParlaMint partner countries and regions, i.e. a response rate of 77%. 26 surveys were fully completed (0.75%) and 9 were partially completed (0.25%), with “partially completed” being those surveys where at least one question was answered.

Of the respondents, just over half (54%) were part of the ParlaMint I phase of the project. Most participants held the following three roles in the project: TEI encoding, preparation and submission of corpus samples, and preparation and submission of the entire corpus (all 17%). Most participants have a background in Natural Language Processing (NLP) (57%), followed by Computer Science (CS) (46%), Digital Humanities (DH), and Humanities (H) (31% each). In addition, three other participants chose “other” background, with the explanation that they come from the fields of Linguistics, Machine Learning, and Physics. It should be noted that the question allowed multiple responses and that there are few participants who have only one particular background. The distribution of participants’ research backgrounds is shown in

¹⁶ParlaMint II also includes corpora for regional (autonomous community) parliaments, namely for Catalonia (ES-CT), Galicia (ES-GA), and Basque Country (ES-PV).

Figure 1 – the most common combination among participants with more than one background was a combination of NLP and Computer Science (9 participants), other combinations (such as NLP and Digital Humanities, NLP and Humanities, or Digital Humanities and Humanities) were less common. In four cases, participants chose a combination of three backgrounds; two cases for a combination of NLP, Computer Science and Digital Humanities; the other two cases for a combination of NLP, Digital Humanities and Humanities.

4.1 TEI

The questionnaire included questions about TEI encoding, participants’ familiarity with TEI before starting the project, their experience with TEI during the project, and their plans to use TEI in the future. In general, 44% of participants were already at least somewhat familiar with the TEI P5 guidelines used in the encoding procedure, followed by 37% of participants who were not familiar with TEI at all. Of these participants, more than half (64%) did not require external sources to become familiar with TEI before they began working on ParlaMint projects (and only followed the guidelines “The structure and encoding of ParlaMint corpora”). In addition, 20% of the participants were very familiar with TEI.

Regarding the experience with TEI prior to ParlaMint, almost half of the participants were already using TEI encoding for their work (48%), while 22% were only using TEI-inspired encoding and 30% had never used TEI. Concerning the experience with TEI encoding in the ParlaMint project, we asked participants to rate various statements about TEI participation in the project and to assess the extent to which they agreed or disagreed with them (rated on a scale of “strongly disagree” (1) to “strongly agree” (5)). Sentiment towards TEI encoding ranged from neutral to mostly positive. Figure 2 shows the results and averaged values of the responses to each individual statement.

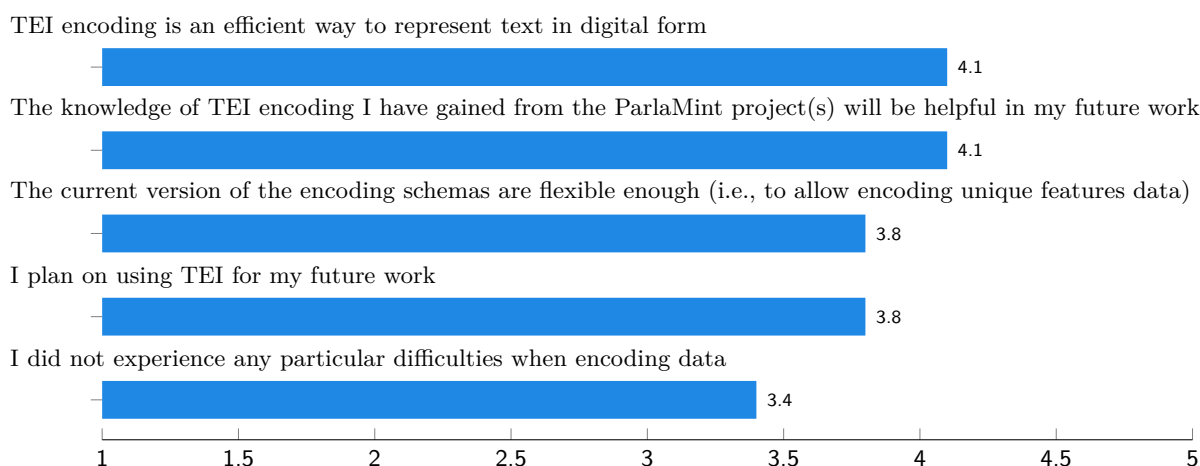


Figure 2: Averaged values of the responses to statements related to the partners’ experience with TEI. The values ranged on the scale of “strongly disagree (1)” to “strongly agree” (5), with (3) indicating a “neutral” position.

The highest rated statement was “*TEI encoding is an efficient way to represent text in digital form*” and “*The knowledge about TEI encoding that I gained from the ParlaMint project will be helpful in my future work*” (both 4.1), followed by “*The current version of the encoding schemes is flexible enough*” and “*I plan to use TEI for my future work*” (both 3.8). The most neutral and lowest rated statement is “*I did not experience any particular difficulties when encoding data*”. This was further explored in the next question, in which we asked participants to provide additional feedback on their experiences with TEI encoding, with some expressing the complexity of TEI encoding in terms of the uniform nature of encoding very different parliamentary systems and languages. For further analysis, we examined the proportion of familiarity with TEI based on participants’ research backgrounds, as shown in Table 1.

Response		Research backgrounds					Total (Distinct count)
		NLP	CS	DH	H	Other	
Very familiar	abs. value	3	1	2	3	1	5
	% of col.	15.00	6.25	18.18	27.27	33.33	14.29
	% of row	60.00	20.00	40.00	60.00	20.00	100.00
Somewhat familiar	abs. value	8	7	5	1	1	12
	% of col.	40.00	43.75	45.45	9.09	33.33	34.29
	% of row	66.67	58.33	41.67	8.33	8.33	100.00
Not at all familiar	abs. value	5	4	2	3	1	10
	% of col.	25.00	25.00	18.18	27.27	33.33	28.57
	% of row	50.00	40.00	20.00	30.00	10.00	100.00
No answer	abs. value	4	4	2	4	0	8
	% of col.	20.00	25.00	18.18	36.36	0.00	22.86
	% of row	50.00	50.00	25.00	50.00	0.00	100.00
Total (responses)	abs. value	20	16	11	11	3	35
	% of col.	100.00	100.00	100.00	100.00	100.00	100.00
	% of row	57.14	45.71	31.43	31.43	8.57	100.00

61

Table 1: Cross-tabulation of the statement “How familiar were you with TEI prior to ParlaMint?” with participants’ research background (35 participants, 61 total responses). Each cell contains the absolute value of participants from each research background and their response, the percentage that the value represents relative to all participants from a given domain, and the percentage that the value represents relative to all participants from each response.

Regarding familiarity with TEI, slightly less than half of the NLP participants were somewhat familiar with TEI, followed by participants who were not familiar with TEI at all, and finally, those that were very familiar with TEI. The distribution trend is similar for both Computer Science and Digital Humanities participants but changes for Humanities where only one participant was somewhat familiar with TEI and the others were either not or very familiar with TEI. Lastly, participants with “other” backgrounds (Linguistics, Physics, and Machine Learning) were evenly distributed (1 participant not at all familiar, 1 participant somewhat familiar, and 1 participant very familiar).

We also checked the responses to the question “The acquired knowledge about TEI will be useful for my future work” in relation to the level of familiarity prior to the ParlaMint project work. Participants who were not at all familiar with TEI at the start of the project were relatively evenly distributed on a range from “strongly agree” to “neutral”. A neutral position was also slightly stronger according to participants who were somewhat familiar with TEI, while the remaining participants voiced agreement (somewhat agree and strongly agree). Finally, more than half of the participants who were very familiar with TEI strongly agreed with the statement.

When asked if they planned to use TEI in their future work (Table 2), NLP participants expressed either strong agreement or a neutral position, and only one expressed mild disagreement. About 30% of Computer Science participants expressed a neutral position, and the others agreed or strongly agreed with the statement. Digital Humanities participants were somewhat mixed, as agreement ranged from “strongly agree” to “neutral,” with a few more choosing “strongly agree.” A similar proportion (and percentages) of agreement that TEI should be useful for their future work was also noted by participants from the Humanities. Finally, participants from “Other” backgrounds were evenly split between “strongly agree,” “somewhat agree,” and “strongly disagree.”

4.2 Git

The questionnaire also included questions about the participants’ familiarity with version control systems (VCS) prior to the start of the ParlaMint project, questions about communication and

Response		Research backgrounds					Total (Distinct count)
		NLP	CS	DH	H	Other	
Strongly agree	abs. value	7	3	4	4	1	10
	% of col.	35.00	18.75	36.36	36.36	33.33	28.57
	% of row	70.00	30.00	40.00	40.00	10.00	100.00
Somewhat agree	abs. value	2	3	2	1	1	5
	% of col.	10.00	18.75	18.18	9.09	33.33	14.29
	% of row	40.00	60.00	40.00	20.00	20.00	100.00
Neutral	abs. value	6	5	2	2	0	10
	% of col.	30.00	31.25	18.18	18.18	0.00	28.57
	% of row	60.00	50.00	20.00	20.00	0.00	100.00
Somewhat disagree	abs. value	1	1	1	0	0	1
	% of col.	5.00	6.25	9.09	0.00	0.00	2.86
	% of row	100.00	100.00	100.00	0.00	0.00	100.00
Strongly disagree	abs. value	0	0	0	0	1	1
	% of col.	0.00	0.00	0.00	0.00	33.33	2.86
	% of row	0.00	0.00	0.00	0.00	100.00	100.00
No answer	abs. value	4	4	2	4	0	8
	% of col.	20.00	25.00	18.18	36.36	0.00	22.86
	% of row	50.00	50.00	25.00	50.00	0.00	100.00
Total (responses)	abs. value	20	16	11	11	3	35
	% of col.	100.00	100.00	100.00	100.00	100.00	100.00
	% of row	57.14	45.71	31.43	31.43	8.57	100.00

61

Table 2: Cross-tabulation of the responses for the statement “I plan to use TEI in future work” on the scale of “Strongly agree” to “Strongly disagree” and participants’ research backgrounds (35 participants, 61 total responses).

workflow, and a general assessment of their experience with Git in the ParlaMint project. A large proportion of participants had previous experience with VCS (28 participants, 80%), whereas 7 participants (20%) had no experience with VCS.

Of the participants who reported previous experience with VCS, when asked which VCS they had used (multiple responses, 27 participants, 56 responses counted), all but one (1 participant did not answer the question) reported experience with Git (GitHub, GitLab: 27, 100%), followed by SVN and Bitbucket (10, 37% each), CVS (6, 22%), and Mercurial (3, 27%). None of the participants reported previous experience with HelixCore or Beanstalk.

Regarding experience with Git prior to starting work on the ParlaMint project, of the participants who indicated previous experience with Git and GitHub in the previous question (27 participants in total), 13 participants (48%) described their experience level as “intermediate”, 9 (33%) as “beginner”, and 5 participants (19%) as “advanced”, which compared to all survey participants (35 in total, 8 or 23% did not provide an answer) equates to 37% “intermediate”, 26% “beginner,” and 14% “advanced”.

Another important aspect of the survey was the communication process, in which the use of GitHub Issues played an important role – GitHub Issues served as the main communication channel between the corpus compilation work-package leads and the project partners, and for communicating issues that partners encountered. They also served to provide information/answers to other project partners facing similar issues. Regarding the communication process with GitHub Issues, participants were mostly positive. Almost 70% used Issues as a means to communicate and discuss problems, and of those, almost all (96%) used Issues to find relevant information pertaining to their particular problem, and more than $\frac{3}{4}$ participated in discussions about problems through Issue comments.

In addition to the use of GitHub Issues, we also asked about the experience with the project

workflow. Most participants (77%) agreed that the Git requirements and workflow were clearly explained, while those who disagreed indicated that some constraints and phases of the workflow were not adequately explained (for example: “the whole process of approving the sample first was unclear to me” and “README was linked to Parla-CLARIN instead of ParlaMint, the upper limit of the sample size was not clearly stated, the paths to the Java libraries were hard-coded, some of the tests assumed Linux (as opposed to OS X), it was unclear that CoNNL-U and vertana were automatically generated”). The same percentage applied to whether they had particular difficulties submitting data samples to GitHub. The difficulties reported ranged from the size of the sample files to the distinction between the process of submitting data samples and the process of submitting full corpora. Over 80% indicated that they received sufficient and good support from the ParlaMint team. When asked if there were other GitHub features that facilitated communication and the work process, several participants mentioned the automatic validation that was started every time they pushed on an open pull request and the helpfulness of the Github Issues (as they provided information for other participants’ issues) or the helpfulness of the people who provided responses to the issues. As a final part of the Git usability assessment, we asked participants to rate their experience with Git, GitHub features, and project workflows in the form of a single multi-sentence response table on a scale of “strongly disagree” (1) to “strongly agree” (5). Figure 3 shows the averages of responses to each statement.

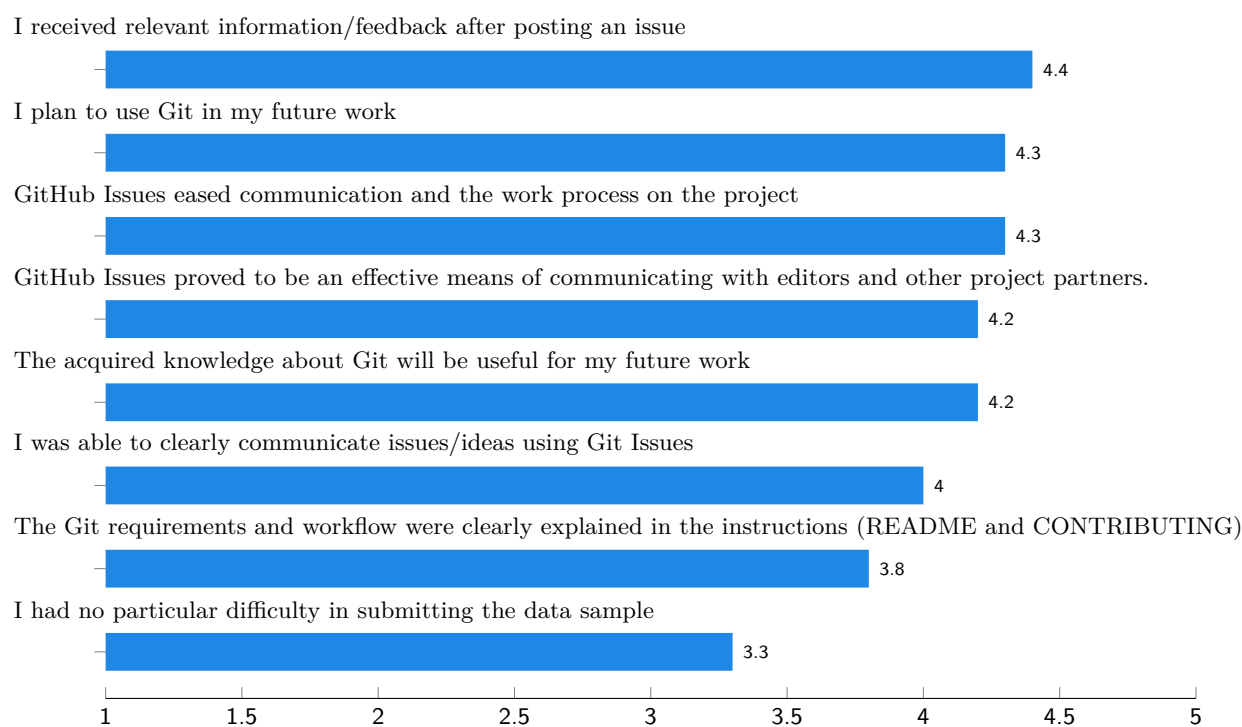


Figure 3: Averaged values of the responses to statements related to the partners’ Git-related experience. The values are ranged on the scale of “strongly disagree (1)” to “strongly agree” (5), with value 3 indicating a “neutral” position.

The highest rated statement was “*I received relevant information/feedback after posting an issue*” (rated 4.4), followed by “*I plan to use Git in my future work*” and “*GitHub issues eased communication and the work process in the project*” (both 4.3). The responses to the statement “*I did not experience any particular difficulties in submitting the data sample*” were neutral (3.3). This was further illustrated by the next question, where we asked for additional comments on the workflow and submission process. There were some comments indicating that new changes to the material in GitHub (e.g., changes to the encoding process and corrections to the validation schemes) were not communicated clearly enough to participants, a video tutorial/workshop was

suggested to facilitate the submission process, and that the GitHub Issues and responses from the team behind them were generally helpful.

We explored several relationships between participants’ responses regarding their (previous) experience with Git and whether they plan to use it in their future work. First, we examined the relationships between participants’ research backgrounds in terms of experience level or familiarity with Git, as shown in Table 3.

Response		Research backgrounds					Total (Distinct count)
		NLP	CS	DH	H	Other	
Advanced	abs. value	2	4	2	0	0	5
	% of col.	10.00	25.00	18.18	0.00	0.00	14.29
	% of row	40.00	80.00	40.00	0.00	0.00	100.00
Intermediate	abs. value	10	6	5	0	3	13
	% of col.	50.00	37.50	45.45	0.00	100.00	37.14
	% of row	76.92	46.15	38.46	0.00	23.08	100.00
Beginner	abs. value	7	5	3	4	0	9
	% of col.	35.00	31.25	27.27	36.36	0.00	25.71
	% of row	77.78	55.56	33.33	44.44	0.00	100.00
No answer	abs. value	1	1	1	7	0	8
	% of col.	5.00	6.25	9.09	63.64	0.00	22.86
	% of row	12.50	12.50	12.50	87.50	0.00	100.00
Total (responses)	abs. value	20	16	11	11	3	35
	% of col.	100.00	100.00	100.00	100.00	100.00	100.00
	% of row	57.14	45.71	31.43	31.43	8.57	100.00

61

Table 3: Cross-tabulation of the question “What was your experience with Git (prior to ParlaMint)” and participants’ research backgrounds.

In NLP, half of the participants described their experience level as “intermediate”, followed closely by beginners and finally two participants at the “advanced” level. In Computer Science, the distribution of experience levels was much more even (5 participants at the “beginner” level, 6 at the “intermediate” level, and 4 at the “advanced” level). Of the participants with a background in the Digital Humanities, slightly less than half described their experience level as “Intermediate”, while in the Humanities, most participants were at the “beginner” level with a large number of “No answer” responses.

Following the above analysis, we also examined the relationship between participants’ opinions about whether the knowledge they gained about Git might prove useful in their future work and their level of experience prior to participating in ParlaMint.

Half of the participants from the beginner group fully agreed that the Git knowledge they acquired would be useful for their future work, while the other half consisted of two participants who tended to agree with the statement and one who slightly disagreed. The sentiment changed in the intermediate group of participants, where more than half strongly agreed with the statement, while almost 30% slightly agreed with the statement and one somewhat disagreed. On the other hand, more than half of the advanced participants held a neutral position.

Finally, we examined the relationship between participants’ responses about their intentions to use Git and their future work (agreement with the statement “I plan to use Git in my future work”) and their research background. The responses are presented in Table 4.

Regarding Git, half of NLP participants expressed strong agreement, while the others only somewhat agreed with the statement or held a neutral position; the same sentiment prevailed among participants from Computer Science and Humanities backgrounds. More than half of all participants with a background in Digital Humanities strongly agreed with the statement, only one somewhat agreed, while participants in the “other” category all expressed strong agreement.

Response		Research backgrounds					Total (Distinct count)
		NLP	CS	DH	H	Other	
Strongly agree	abs. value	10	6	7	3	3	16
	% of col.	50.00	37.50	63.64	27.27	100.00	45.71
	% of row	62.50	37.50	43.75	18.75	18.75	100.00
Somewhat agree	abs. value	3	3	1	1	0	5
	% of col.	15.00	18.75	9.09	9.09	0.00	14.29
	% of row	60.00	60.00	20.00	20.00	0.00	100.00
Neutral	abs. value	2	3	0	1	0	4
	% of col.	10.00	18.75	0.00	9.09	0.00	11.43
	% of row	50.00	75.00	0.00	25.00	0.00	100.00
Strongly disagree	abs. value	1	0	0	1	0	1
	% of col.	5.00	0.00	0.00	9.09	0.00	2.86
	% of row	100.00	0.00	0.00	100.00	0.00	100.00
No answer	abs. value	4	4	3	5	0	9
	% of col.	20.00	25.00	27.27	45.45	0.00	25.71
	% of row	44.44	44.44	33.33	55.56	0.00	100.00
Total (responses)	abs. value	20	16	11	11	3	35
	% of col.	100.00	100.00	100.00	100.00	100.00	100.00
	% of row	57.14	45.71	31.43	31.43	8.57	100.00

61

Table 4: Cross-tabulation of the statement “I plan to use Git in my future work” and participants’ research backgrounds.

5 Conclusions

The paper attempted to show how TEI can be used to specify the encoding of complex language corpora (or other types of language resources), providing both the guidelines of those wishing to encode the corpora, as well as XML schemas that are used to formally validate their encoding. We also presented Git which is well suited for controlled and distributed development and publishing of not only the guidelines and schemas but also the language resources themselves. As mentioned, the size of the produced ParlaMint corpora makes it somewhat problematic to store them in their entirety on GitHub but this is not the case for smaller language resources, particularly manually annotated ones.

Fully mastering Git is also not a simple process, especially for the typical researcher of the target Social Sciences (SSH) community. However, with an appropriate set-up, such as we have attempted to provide for ParlaMint, we believe that with only basic knowledge, partners can successfully submit their data sample with an automatic check if they validate, while the validation of the complete corpus still relies on local processing.

The paper also presented a survey on the use of TEI and Git by the ParlaMint project partners. Overall, participants were mostly positive about their experience with TEI and Git, although some difficulties were reported, mainly related to the distinction between the submission process for the sample corpus and the full corpus, as well as some workflow limitations. The difficulties identified will serve as an opportunity to update and optimise the current project workflow. With respect to TEI, initial responses were mixed. There was agreement that TEI is an efficient way to represent text in digital form and that the lessons learned will help participants in their future work. Contrary, there was less agreement, or even neutrality, on the question of whether the current schema is flexible enough to support the encoding of unique features of the data, given that it is still a very uniform and sometimes complex way of encoding sometimes drastically different parliamentary systems. On the other hand, reactions to Git were very positive, from relevant information and feedback received via GitHub Issues, to effectiveness in the communication process, to plans to use Git in the future. There was less agreement on whether the

requirements and workflow were adequately explained, which points to the difficulties in data submission mentioned above.

The survey helped us to gain some insights into the relationships between participants' research backgrounds, their level of experience and opinion of their TEI and Git experience in the ParlaMint project, and participants' opinions on whether they plan on integrating them into their future work. We recognise that given the small number of participants (though still reasonably representative of the ParlaMint project group, as there are not many project participants), firm or decisive conclusions can not be drawn. Nonetheless, the results of the cross-tabulations can give us some insight into the role that research background and project work play in the likelihood that project partners would include Git and TEI in their "digital toolbox."

The survey showed that participants with Digital Humanities, NLP, and especially Humanities backgrounds tended to be more familiar with TEI. Among participants who had less or no prior familiarity with TEI, there was high agreement that the knowledge they gained would be helpful in their future work.

In comparison, as expected, the responses on the use of Git showed that participants from the NLP and Computer Science fields already relied on and used it regularly. In contrast, participants from the Digital Humanities and especially the Humanities were either unfamiliar or not very familiar with Git at the beginning of the project work. Regarding prior familiarity with Git and agreement on whether the knowledge gained could help them in their future work, most participants agreed, with the exception of those who were already very familiar with Git, most of whom were neutral about it, likely due to the fact that the project did not provide them with any new experiences they had not had before. Finally, most participants agreed that they would like to use Git for their work in the future, and this sentiment was particularly evident in the areas of Digital Humanities and NLP.

In our future work, we plan to continue working on the ParlaMint ODD, which will be updated as we move on to new types of annotation, in particular semantic annotation, and new types of resources, in particular machine-translated corpora and speech data.

The development of the ParlaMint corpora is also currently still rather centralised. In the longer perspective, we would like to encourage anyone that would wish to produce a ParlaMint-compatible corpus to be able to do so independently, for which we have to make the set-up (even) more flexible and also provide a tutorial on how to independently produce a ParlaMint compatible corpus. It is important to note that the corpus generated and validated with the automatic validation will be ParlaMint compatible only at the schema level, as, at least so far, we have also done manual verification in order to ensure the quality of the corpus data.

We believe that both TEI and especially Git, or, in principle, some other Version Control System – and the possibilities of combining the two – are not as well known in the SSH community as they should be, and that learning about them and adopting them into the work process could go a long way in making the development of encoding guidelines and language resources a much smoother and more controlled process, also leading to better reproducibility, a point that is very relevant to the goals of the CLARIN infrastructure.

Acknowledgements

The authors would like to thank the three anonymous reviewers for their helpful comments and suggestions, and all the partners in ParlaMint I and II for their suggestions for improving the ParlaMint schema and tools and for taking part in the survey. The work presented in this paper was supported by the CLARIN ERIC projects ParlaFormat (2019) and "ParlaMint: Towards Comparable Parliamentary Corpora" (2020–2021 and 2022–2023), by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ, and individual project partner grants.

References

- Lou Burnard, Christof Schöch, and Carolin Odebrecht. 2021. In Search of Comity: TEI for Distant Reading. *Journal of the Text Encoding Initiative*, (14).
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 07.
- Tomaž Erjavec and Andrej Pančur. 2021. The Parla-CLARIN Recommendations for Encoding Corpora of Parliamentary Proceedings. *Journal of the Text Encoding Initiative*. <https://doi.org/10.4000/jtei.4133>.
- Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michal Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Dargis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2022. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*. <https://doi.org/10.1007/s10579-021-09574-0>.
- Christof Schöch, Roxana Patraş, Diana Santos, and Tomaž Erjavec. 2021. Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*, (1). <http://doi.org/10.3828/mlo.v0i0.364>.
- Toma Tasovac, Laurent Romary, Piotr Banski, Jack Bowers, Jesse de Does, Katrien Depuydt, Tomaž Erjavec, Alexander Geyken, Axel Herold, Vera Hildenbrandt, Mohamed Khemakhem, Boris Lehečka, Snežana Petrović, Ana Salgado, and Andreas Witt. 2018. TEI Lex-0: A baseline encoding for lexicographic data. Version 0.9.1. Technical report, DARIAH Working Group on Lexical Resources.
- TEI Consortium. 2022. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium. <http://www.tei-c.org/Guidelines/P5/>.

EU Data Governance Act: Outlining a Potential Role for CLARIN

Paweł Kamocki
IDS Mannheim,
Germany
kamocki@ids-
mannheim.de

Krister Lindén
University of Helsinki,
Finland
krister.linden@
helsinki.fi

Andrius Puksas
Mykolas Romeris
University,
Lithuania
andrius_puksas@
mruni.eu

Aleksei Kelli
University of
Tartu,
Estonia
aleksei.kelli@
ut.ee

Abstract

The Data Governance Act was proposed in late 2020 as part of the European Strategy for Data, and adopted on 30 May 2022 (as Regulation 2022/868). It will enter into application on 24 September 2023. The Data governance Act is a major development in the legal framework affecting CLARIN and the whole language community. With its new rules on the re-use of data held by the public sector bodies and on the provision of data sharing services, and especially its encouragement of data altruism, the Data Governance Act creates new opportunities and new challenges for CLARIN ERIC. This paper analyses the provisions of the Data Governance Act, and aims at initiating the debate on how they will impact CLARIN and the whole language community.

1 Introduction

The third decade of the 21st century has started with some of the most perturbing events in generations: the COVID-19 pandemic and the Russian aggression against Ukraine, which – understandably so – overshadowed all other developments. Meanwhile, however, the European Union is dynamically modernising its legal framework concerning digital data, or, as the European Commission itself would put it, “shaping Europe’s digital future”¹.

In recent years, the adoption of the General Data Protection Regulation (GDPR) and its entry into application have shown that the European Union is a true global “regulatory superpower”. As one of the world’s largest markets (also in terms of population, but especially in terms of purchasing power), the EU has the power to influence the policies of manufacturers of goods and providers of services worldwide, by imposing standards on goods and services that can enter its market. With the so-called “Brussels effect” (Bradford, 2020), the GDPR has become the global standard for personal data protection. This phenomenon is now likely to extend to other types of digital data, markets, and services. This is why the developments in the EU’s legal framework affecting this domain are closely followed not only in Europe, but also on other continents, as they have the potential to – literally – change the digital world.

The Data Governance Act (DGA) can be considered one step further in data governance. The DGA aims to reduce the digital divide, ensure data access neutrality, portability and interoperability, and avoid lock-in effects (Recital 2 of DGA). Its objective is also to improve the conditions for data sharing in the internal market by creating a harmonised framework for data exchanges and laying down certain basic requirements for data governance, paying specific attention to facilitating cooperation between Member

¹ <https://digital-strategy.ec.europa.eu/en/policies/strategy-data>

States“ (Recital 3 of DGA). These aims are underlined by the idea that “data that has been generated or collected by public sector bodies or other entities at the expense of public budgets should benefit society” (Recital 6 of DGA).

This paper will present the content of the DGA from the perspective of CLARIN ERIC and the EU language community as a whole and discuss the impact that this Regulation may have on the functioning of CLARIN ERIC in the future.

In the second section, the authors analyse the concept of data covered by DGA and its interaction with personal data and intellectual property protection. The third section explains the process that led to the adoption of the DGA, and the fourth section explains its contents.

2 Rights in Re-usable Data

Article 2 (1) of the Data Governance Act defines data as “any digital representation of acts, facts or information and any compilation of such acts, facts or information, including in the form of sound, visual or audiovisual recording”. The definition of data in DGA is wide and potentially encompasses different types of material. It is similar to definitions of research and language data and therefore entails similar challenges. For further discussion, see Kelli et al. (2020); Kelli et al. (2018).

Recital 12 of the Data Governance Act explains that the re-use² regime applies to data that “*the supply of which forms part of the public tasks of the public sector bodies*”. The Chapter of the DGA on the re-use of data provides that it applies to data protected on grounds of statistical confidentiality, intellectual property right (IPR) protection and personal data (Article 3 (1)). Issues related to re-use of data containing IP and personal data are discussed below.

Data defined by the DGA could be subject to intellectual property rights (IPR). According to Recital 17 of the Data Governance Act “This Regulation should neither affect the existence or ownership of intellectual property rights of public sector bodies nor limit the exercise of those rights in any way”. At the same time, Recitals 17 and 18 of the DGA require public sector bodies to exercise their IPRs in a way that facilitates re-use.

The DGA also has more detailed instructions on dealing with IPRs within the framework of conditions for re-use. Article 5 (3) (a) (ii) provides that public sector bodies grant access for the re-use of data only where it has ensured that data has been “modified, aggregated or treated by any other method of disclosure control, in the case of commercially confidential information, including trade secrets or content protected by intellectual property rights.” At the same time, Article 5 (7) of the DGA prescribes that reuse must be in compliance with intellectual property rights.³

Suppose the described methods (i.e. modifying, aggregating) are applied to trade secrets and confidential information. In that case, it should not pose any other legal problems than to need a guarantee that secret information is not revealed. When it comes to IP-protected content (especially copyright-protected works), the situation is different. For further discussion on copyright-protected works as language data, see Kelli et al. (2022).

The challenge with modifying copyright-protected works relates to the existence of moral rights (especially the right of integrity). According to Article 6*bis* of the Berne Convention, the author has the right “to object to any distortion, mutilation or other modification of, or other derogatory action in relation to, the said work, which would be prejudicial to his honor or reputation”. Therefore, to avoid copyright violation, it is essential not to distort or mutilate works during the modification and aggregation process.

There are also further requirements for transmitting data protected by intellectual property rights to a re-user who intends to transfer those data to a third country (Article 5 (10)).

The DGA divides data into two main groups: personal data and non-personal data (Article 2 (3), (4)). According to Article 4 (1) of the General Data Protection Regulation, personal data means “any information relating to an identified or identifiable natural person (‘data subject’)”. Recital 4 of the Data

² Article 2 (2) of the DGA defines ‘re-use’ as “the use by natural or legal persons of data held by public sector bodies, for commercial or non-commercial purposes other than the initial purpose within the public task for which the data were produced, except for the exchange of data between public sector bodies purely in pursuit of their public tasks”.

³ Recital 10 of the DGA emphasizes the same principle for trade secrets: “The re-use of data, which may contain trade secrets, should take place without prejudice to Directive (EU) 2016/943, which sets out the framework for the lawful acquisition, use or disclosure of trade secrets”.

Governance Act emphasises that “this Regulation should not be read as creating a new legal basis for the processing of personal data for any of the regulated activities [...] In the event of a conflict between this Regulation and Union law on the protection of personal data or national law adopted in accordance with such Union law, the relevant Union or national law on the protection of personal data should prevail”. This means that processing personal data is regulated by the General Data Protection Regulation and not by the Data Governance Act. Recital 7 of the DGA repeats that legal bases for the processing of personal data is provided in the GDPR.

However, the DGA still foresees re-use of data containing personal data. Article 3 (1) (d) of the Data Governance Act *expressis verbis* provides that the re-use regulation “applies to data held by public sector bodies which are protected on grounds of the protection of personal data”. At the same time, Article 5 (3) (a) (i) of the DGA obliges public sector bodies to ensure protection of personal data. Therefore they may grant access or the re-use of data only where public sector bodies have anonymised personal data.

Specific conditions for re-use of data containing personal data or third-party IPRs are discussed below.

3 Toward the Data Governance Act

On 19 February 2020, just a couple of weeks before the first COVID lockdowns, the European Commission launched the European Strategy for Data (European Commission, 2020a). This was followed by a large stakeholders consultation (lasting until 31 May 2020), in which 806 contributions were received, including 98 from academic/research institutions.⁴ A series of proposals for Regulations (labelled, in the Anglo-Saxon way, “Acts”) were adopted based on this consultation, including:

- Data Governance Act (25 November 2020);
- Digital Services Act (15 December 2020);
- Digital Markets Act (15 December 2020);
- Artificial Intelligence Act (21 April 2021);
- Data Act (23 February 2022).

So far (as of February 2023), the first three of these Acts were adopted: apart from the Data Governance Act (which, after a few modifications, has become the Regulation 2022/868 of 30 May 2022), also the Digital Services Act (Regulation (EU) 2022/2065 of 19 October 2022) and the Digital Markets Act (Regulation (EU) 2022/1925 of 14 September 2022). Both the Artificial Intelligence Act and the Data Act are in advanced stages of their legislative processes.

The DGA will enter into force on 24 September 2023.

4 The Content of the Data Governance Act

As expressly stated in the explanatory memorandum accompanying the Commission’s proposal for the DGA (European Commission, 2020b), the Act draws inspiration from the principles for data management and re-use developed for research data, namely the FAIR data principles stipulating that data should, in principle, be findable, accessible, interoperable and re-usable. This principle with a particular attention to a high level of cybersecurity is emphasised in Recital 2 of DGA as well. The Act’s announced aim is to address a number of rather diverse practical issues, i.e.:

Making public sector data available for reuse in situations where such data is subject to rights of others.

- Sharing of data among businesses, against remuneration in any form.

⁴ https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12271-European-Strategy-for-data/public-consultation_en

- Allowing personal data to be used with the help of a ‘personal data-sharing intermediary’, designed to help individuals exercise their rights under the General Data Protection Regulation (GDPR).
- Allowing data use on altruistic grounds.

From the perspective of the language community, the most important changes introduced by the DGA can be summarised in four points:

- The Act enables wider re-use of ‘protected’ data held by public sector bodies;
- The Act introduces a mechanism for supervising providers of data sharing services;
- The Act promotes a new paradigm in sharing data for purposes of general interest, referred to as ‘data altruism’.
- The Act establishes a European Data Innovation Board.

Each of these points is developed in a relevant sub-section below.

4.1 Wider Re-usability of ‘Protected’ Data Held by Public Sector Bodies

Public sector bodies (i.e., according to the legal definition [Art. 2(1) of the Open Data Directive], States, regional and local authorities, and bodies established for general interest purposes and financed by the State or regional and local authorities) produce and collect massive amounts of data, which represent considerable economic value. Since the collection and/or production of such data is indirectly financed by the taxpayers, it is justifiable to make these data available for anyone, and re-usable for any purpose, be it commercial or not. This was the general logic behind the Public Sector Information (PSI) Directive (2003/98/EU, amended by Directive 2013/37/EU), which was recently amended by the Open Data Directive (EU) 2019/1024.

Under this framework, Public Sector Information seemed an interesting source of research data for the language community; it was used as one of the main sources of data in the European Language Resources Coordination (ELRC) initiative (Lösch et al., 2018), aimed at collecting Language Resources for the development of automated translation solutions. The amount of public sector data collected within the ELRC initiative (in which Kamocki, one of the co-authors of this paper, was temporarily involved as a legal expert), was rather significantly limited by the fact that a large portion of language data, mono- or multilingual, is effectively excluded from the scope of the rules on the re-use of PSI. Most importantly, under PSI/Open Data Directive documents in which third parties (i.e. anyone else than the public body holding the document) have intellectual property rights, as well as documents containing personal data, are not considered reusable (cf. Article 1(2)(c) and (h) of the Open Data Directive). For example, if a translator (or a translation provider) holds copyright in a document, this document is not covered by the rules on re-use of public sector information under the Open Data Directive. The DGA aims at mitigating this situation.

Under the DGA, public sector bodies have a general obligation to make their datasets available for re-use, even if they contain personal data or data protected by third-party copyright. It does not mean, however, that copyright or personal data protection no longer applies to documents held by public sector bodies; rather, these bodies are obliged to find solutions that enable re-use of concerned datasets in a way that respects the rights of third parties. This can be achieved by sharing pre-processed data (e.g., anonymised, aggregated or otherwise modified, possibly also in a ‘derived text format’, cf. Article 5(3)(a) of the DGA). Another possibility is to limit access to the data to a secure processing environment provided or controlled by the public sector body (Article 5(3)(b) and (c) of the DGA). Such secure processing environments are already in use for the processing for research purposes of statistical microdata held by Eurostat (on the basis of EU Regulation No 557/2013 No 557/2013), or in the Leibniz Institute for the Social Sciences, also for the processing of statistical microdata (Kamocki et al., 2016). It is to be expected that most public sector bodies will not be able to provide sophisticated virtual data rooms, such as those used in some merger and acquisition transactions, but a simple dedicated room with no Internet access, which cannot be entered with any electronic device, can probably be a sufficient

solution in most cases. Finally, if providing access to the data in secure conditions proves impossible, the public sector body should “make best efforts” to help potential re-users obtain necessary consents or permissions from rightholders (Article 5(7) of the DGA).

For making data available for reuse, public sector bodies may charge fees, but these fees should not exceed the costs incurred; they should also be transparent and non-discriminatory (Article 6 DGA). Moreover, public sector bodies may apply reduced fees, or even allow re-use free of charge, for certain categories of re-users, such as start-ups, small and medium-sized enterprises (SMEs) and educational establishments.

This may sound extremely promising in terms of opening vast amounts of data for re-use; however, the direct impact of the new rules on the research community will be limited due to the fact that, unlike the Open Data Directive, the relevant provisions of the DGA do not include data held by educational and cultural establishments in its scope (Article 3(2)(c) of the DGA). This means that ‘protected’ data held by museums, archives or libraries are not concerned by the abovementioned rules, and are not to be made available for reuse. The situation of ‘protected’ research data (for example, corpora of copyright-protected language data, or speech recordings) held by universities is in fact quite uncertain. On the one hand, universities qualify as ‘education establishments’, and as such should be excluded from the relevant provisions of the DGA. Recital 12 of the DGA, however, suggests the contrary; it reads:

“(…) Research-performing organisations and research-funding organisations could also be organised as public sector bodies or bodies governed by public law.

This Regulation should apply to such hybrid organisations only in their capacity as research-performing organisations.”

This seems to indicate that research data held by universities are in fact concerned by the rules on re-use, which has the potential of transforming the role of CLARIN and its centres as providers of (mostly ‘protected’) research data. One could imagine that in a world where universities and public research organisations are obliged by law to make their research data available for re-use, CLARIN could effectively become an essential intermediary, a one-stop-shop for researchers EU-wide, where they can at least learn which dataset is available at what place, and what are its re-use conditions. Such an intermediary, referred to as “sectoral information point”, is mentioned in Article 8 of the DGA, the Article providing that Single Information Points shall be established in every Member State, and a Single European Information Point should be established by the Commission. These Single Information Points shall make available a searchable list of datasets available for re-use in the relevant area (together with information on their size and conditions for re-use). Some of these datasets can be available through Sectoral Information Points. Information Points should work as interfaces for re-users – they shall receive requests for re-use of datasets which are present on their lists, and transmit these requests to the public sector bodies that hold the data. In this approach, a CLARIN centre could become a Sectoral Information Point for Language Data, and maintain a list of all the relevant datasets available in their area (held by universities, but also, e.g., by ministries or local authorities). A user would consult the list, file a request for re-use with the centre, which will then transmit it to the body (e.g., a university) who holds the data, in order to enable the user to access the data through the body’s own secure processing environment. This role of a Sectoral Information Point could complement the data hosting task.

Moreover, in order to assist public sector bodies in fulfilling their obligations under the DGA (including to provide them with technical support, e.g. in the field of data anonymisation), Member States should create ‘competent bodies’ with adequate legal and technical capabilities and expertise (Article 7 of the DGA). This can be done either by establishing new bodies, or entrusting existing bodies (like, for example, National Data Protection Authorities) with these missions. This has a twofold benefit for CLARIN – firstly, CLARIN centres could share their expertise in the field of data sharing with those competent bodies, thereby contributing to making (language) data more widely accessible while also increasing the visibility of the CLARIN network. Secondly, these competent bodies will likely issue guidelines and recommendations regarding sharing of protected data, which will shed light on some of the grey areas of the field, and facilitate CLARIN’s own mission.

To sum up, the provisions of the DGA related to the re-use of ‘protected’ data held by public sector bodies (Chapter II) create many opportunities for CLARIN and the EU language community as a whole – first of all, a wealth of new data, including language data, will be made available for reuse, including for such purposes as developing language resources, and training language models. Secondly, CLARIN

could act as a Sectoral Information Point, thereby assisting researchers in accessing datasets that cannot be hosted in a CLARIN centre due to copyright- or data-protection-related constraints. Thirdly, the DGA will stimulate the development of anonymisation and pseudonymisation techniques and standards, as well as secure solutions for data sharing, which will also open new possibilities for language data. Fourthly and finally, CLARIN's expertise in processing language data can potentially be interesting for the 'competent bodies', which will create new possibilities for liaising with other actors of the data economy, and increase the visibility of CLARIN and its activities.

4.2 New Framework for Data Intermediation Services

Chapter III of the DGA contains some requirements applicable to data intermediation services. These services are defined to include, among others (Article 10 of the DGA):

intermediation services between data holders and potential data users, including making available the technical or other means to enable such services; those services may include bilateral or multilateral exchanges of data or the creation of platforms or databases enabling the exchange or joint use of data, as well as the establishment of other specific infrastructure for the interconnection of data holders with data users.

The activities of CLARIN ERIC (and/or CLARIN B-centres) undoubtedly fall within the scope of this definition. However, it seems that in principle CLARIN and its centres are covered by the exception of Article 15 of the DGA, according to which:

[the relevant provisions] shall not apply to recognised data altruism organisations (see below – authors) or other not-for-profit entities insofar as their activities consist of seeking to collect data for objectives of general interest, made available by natural or legal persons on the basis of data altruism, unless those organisations and entities aim to establish commercial relationships between an undetermined number of data subjects and data holders on the one hand and data users on the other.

This suggests that Chapter III of the DGA applies only to those CLARIN centres which decided – or will decide in the future – to get involved in commercial relationships ('with an undetermined number of partners', i.e., probably, making data commercially available to anyone interested). Since such a turn may be considered by some centres, or even the ERIC as a whole, it is worthwhile to discuss here the impact of the DGA on commercial data sharing.

Under Articles 10 and 11, the provision of 'data sharing services' is subject to notification (Articles 10 and 11). This notification is to be made to a 'competent authority' (such authorities are to be designated by every Member State), whose task is to monitor compliance with a set of obligations listed in Article 12 of the DGA. Those include, e.g.,

- the prohibition of re-using data for other purposes than providing them to the users and the obligation to place data sharing services under a separate legal entity (Article 12(a));
- the prohibition to use metadata collected from the provision of the service for other purposes than developing the service (Article 12(c));
- the general obligation to keep the data in the format in which they were provided by the user (Article 12(d)) and
- the obligation to ensure continuity of service and access to the data in case of insolvency (Article 12(h)).

Failing to meet these conditions may result in 'dissuasive' financial penalties and/or cessation of the provision of the service (Article 14(4)).

This strict framework, intended to instil trust in data sharing within the European Data Spaces, may in fact dissuade CLARIN and its centres from involvement in commercial data sharing.

4.3 Promotion of Data Altruism

It has been noted that many individuals and companies are willing to 'donate' their data for a general interest purpose. This phenomenon is generally observed in the domain of medical research, but it can also affect other areas of research: for example, one can imagine writers or publishers who agree to have

their books and other publications analysed by scientists “for the greater good”. Chapter IV of the DGA introduces a legal framework for such ‘data donations’, which it calls ‘data altruism’.

‘Data altruism’ is defined as ‘the voluntary sharing of data on the basis of the consent of data subjects to process personal data pertaining to them, or permissions of data holders to allow the use of their non-personal data without seeking or receiving a reward (...) where they make their data available for objectives of general interest (...) [such as] scientific research purposes (...)’ (Article 2(16) of the DGA).

The data can be ‘donated’ by signing a European data altruism consent form (Article 25 of the DGA), which will be adopted by the European Commission at a later date.

The key role in this new framework is played by so-called ‘certified data altruism organisations’, trusted organisations that function as intermediaries between the data donors and the users. Their task is to ensure that the data are only used for the purposes of general interest for which they were ‘donated’.

It may seem that CLARIN could consider becoming such a ‘certified data altruism organisation’, although it is not an easy task to accomplish. In order to be certified, an organisation needs to meet objectives of general interest and operate on a non-for profit basis, independently from any for-profit entities (such as private research companies), and through a legally independent structure (Article 18 of the DGA). Moreover, it needs to respect a Rulebook which will be adopted by the European Commission at a later date (Article 22 of the DGA); the Rulebook will specify such aspects as technical requirements related to the storage of the data (regarding security and interoperability standards), or information to be provided to data donors.

Data altruism organisations which meet these requirements shall be subject to registration with a competent authority (which are to be designated by each Member State); they are bound to respect transparency requirements *vis-à-vis* the data holders (Article 20(1) of the DGA; e.g., a full up-to-date list of entities granted access to the data should be provided together with the purpose of processing declared by each of those entities). Furthermore, these organisations will be obliged to submit an annual activity report to the national competent authority (Article 20(2)); if applicable the report shall include a summary of results of the data use allowed by the organisation.

The mechanism of ‘data altruism consent’ could of course be extremely beneficial – it could potentially make most important legal hurdles in access to language data go away – but the status of a ‘data altruism organisation’ would require changes in how CLARIN centres operate. In particular, the data obtained through data altruism cannot be made available for re-use to everyone and for every purpose (i.e., under ‘open’ conditions), but only to researchers, possibly with an obligation to report back on the results. It remains unclear if a ‘data altruism organisation’ can also provide access to data obtained through other means than the ‘data altruism consent’, e.g. via open licences or directly from the public domain. Finally, it is not clear whether ERICs are to be granted any form of special treatment with regards to this aspect of the DGA (such as automatic recognition as ‘data altruism organisations’). It could be envisaged to split the responsibilities between the various CLARIN entities – some of them (e.g. one per consortium,) could indeed become certified data altruism organisations, while others could operate with commercial actors as ‘regular’ data intermediaries (see above). It is also envisageable to assign the role of a certified data altruism organisation only to CLARIN ERIC, which will then grant access to the data to the CLARIN centres.

In sum, for CLARIN ERIC, ‘data altruism’ is a central element of DGA, one which will certainly have to be discussed at the highest level. Unfortunately, as for today many elements of this framework remain unclear – but it is not too early to start a debate.

4.4 Establishment of the European Data Innovation Board

Last but not least, the DGA also provides that the European Commission shall establish a European Data Innovation Board (Chapter VI of the DGA). It will consist of representatives of national competent authorities for data intermediation (see above) and for data altruism (see above), as well as, inter alia, the European Data Protection Board and the European Union Agency for Cybersecurity. In addition, the Commission will launch a call for experts to invite additional members.

The European Data Innovation Board will be divided into at least three sub-groups, one of which will be expressly devoted to standardisation, portability and interoperability of data. Apart from acting as an

advisor to the Commission, the Board will also issue guidelines for Common European Data Spaces, addressing such issues as data interoperability and cross-sectoral standardisation.

In the future, CLARIN will certainly benefit from the guidelines issued by the European Data Innovation Board, or even contribute, directly or indirectly, to their adoption.

5 Conclusion

The Data Governance Act, after its entry into application in September 2023, will be a major development in the legal framework affecting CLARIN and the whole language community. With its new rules on the re-use of data held by the public sector bodies and on the provision of data sharing services, and especially its encouragement of data altruism, the Data Governance Act creates new opportunities and new challenges for CLARIN ERIC.

References

- Berne Convention. Berne Convention for the Protection of Literary and Artistic Works of September 9, 1886. Available at <https://wipolex.wipo.int/en/text/283698> (26.1.2023).
- Bradford, A. 2020. *The Brussels Effect: How the European Union Rules the World*. Oxford University Press, USA.
- DGA. Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act). OJ L 152, 3.6.2022, p. 1–44. Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868&qid=1674669989345> (25.1.2023).
- European Commission. 2020a. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of Regions. A European strategy for data*. COM/2020/66 final.
- European Commission. 2020b. *Proposal for a Regulation of the European Parliament and of the Council on European Data Governance (Data Governance Act)*. COM(2020) 767 final.
- GDPR. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, p. 1-88. Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555312258399&uri=CELEX:32016R0679> (25.1.2023).
- Kamocki, Paweł, Katharina Kinder-Kurlanda, Marc Kupietz (2016). „One Does Not Simply Share Data“. Organisational and Technical Remedies to Legal Constraints in Research Data Sharing – building bridges between Digital Humanities and the Social Sciences. In: When DH Meets Law: Problems, Solutions, Perspectives. Multiple Paper Session at Digital Humanities 2016, Krakow, Poland, 12-16 July 2016.
- Kelli, Aleksei, Arvi Tavast, Krister Lindén. 2022. Building a Chatbot: Challenges under Copyright and Data Protection Law. In: Martin Ebers, Cristina Poncibò, Mimi Zou (Ed.). *Contracting and Contract Law in the Age of Artificial Intelligence*. (115–134). Hart Publishing. DOI: <http://dx.doi.org/10.5040/9781509950713.ch-007> (26.1.2023).
- Kelli, Aleksei, Arvi Tavast, Krister Lindén, Kadri Vider, Ramunas Birštonas, Penny Labropoulou, Irene Kull, Gaabriel Tavits, Age Väriv, Pavel Stranák, Jan Hajic. 2020. The Impact of Copyright and Personal Data Laws on the Creation and Use of Models for Language Technologies. In: Kiril Simov, Maria Eskevich (Ed.). *Selected Papers from the CLARIN Annual Conference 2019* (53–65). Linköping University Electronic Press. Available at <https://doi.org/10.3384/ecp2020172008> (25.1.2023).
- Kelli, Aleksei, Tõnis Mets, Lars Jonsson, Krister Lindén, Kadri Vider, Ramūnas Birštonas, Age Väriv. 2018. Challenges of Transformation of Research Data into Open Data: the Perspective of Social Sciences and Humanities. *International Journal of Technology Management and Sustainable Development*, 17 (3), 227–251.
- Lösch, Andrea, Valérie Mapelli, Stelios Piperidis, Andrejs Vasiļjevs, Lilli Smal, Thierry Declerck, Eileen Schnur, Khalid Choukri and Josef van Genabith. 2018. European Language Resource Coordination: Collecting Language Resources for Public Sector Multilingual Information Management. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 1339-1343.

Open Data Directive. Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast). - ELT L 172, 26.6.2019, p. 56-83.

Semantic Classification of Prepositions in BulTreeBank WordNet

Zara Kancheva

Artificial Intelligence and Language Technologies Department,
Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences, Bulgaria
zara@bultreebank.org

Abstract

The paper presents the work in progress for a PhD thesis about preposition incorporation in the Bulgarian BulTreeBank WordNet. Being one of the most polysemous parts of speech, prepositions are still relatively challenging for NLP and are usually missing in wordnets. A preposition semantic classification, a model for preposition synsets and synset relations are proposed. The planned applications of the prepositions and the directions for future processing are introduced.

1 Introduction

The aim of the paper is to present the work in progress for a thesis about incorporation of prepositions in the structure of BulTreeBank WordNet (BTB-WN) for Bulgarian. Prepositions are viewed as a necessary part of speech in a lexical resource like wordnet, because their integration would seriously expand its range of applications and would be beneficial for several NLP tasks (such as semantic annotation, word sense disambiguation, machine translation, parsing, knowledge extraction, word embeddings, text analysis and generation, etc.). Prepositions are governing words and present the language-specific semantics more completely, so they would contribute to better automatic translation generation and better word embeddings in Bulgarian. The processing and representation of prepositions is a challenging task mainly due to their high polysemy, so they are often missing in lexical resources.

In addition to the extension of BTB-WN the next stage of the work is going to be dedicated to neural models building and prepositions are important for this task because they have a considerable role in the semantics of the text. Algorithms will be developed for the generation of artificial texts in Bulgarian and these texts will use the semantic classification of prepositions and their relations with BTB-WN in order to produce more natural pseudo corpora. Within CLaDA-BG¹ the work will play manifold roles: 1) representation of preposition semantics; 2) supporting language model training for Bulgarian; 3) NLP applications such as word sense disambiguation. The following features will be considered: semantic classification of prepositions, the BTB-WN categories of verbs and nouns to which the prepositions refer and later the frames of the verbs and the semantic roles of the nouns from the Bulgarian OntoValence lexicon. A model for preposition synsets and relations in BTB-WN will be presented.

Explanatory dictionaries typically contain prepositions, but do not present them in a way that is satisfactory for NLP tasks. Some resources such as treebanks, parsers, etc. include prepositions, but for wordnets this part of speech is unusual. The processing of prepositions typically includes semantic and syntactic classification, thematic roles, categorisation by wordnet hierarchy structure or data from resources like valency lexicons, FrameNet², VerbNet³, PennTreebank (Marcus et al., 1994) and PropBank⁴.

In this study a semantic classification of Bulgarian prepositions is done and the classes are used as synset categories for the prepositions, like the synsets for any other part of speech have such categories.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

¹<https://clada-bg.eu/en/>

²<https://framenet.icsi.berkeley.edu/fndrupal/>

³<https://verbs.colorado.edu/verbnet/>

⁴<https://propbank.github.io/>

Preposition synsets have the same structure and features as the synsets for other parts of speech in BTB-WN, however less relations are relevant for them. These synsets have definitions, examples, synonyms and antonyms if available, and the synset category shows their semantic class. This topic is going to be further explored in the future work, because some studies (Amaro, 2018; Harabagiu, 1996) show that there are more applicable relations for preposition synsets, for example *hyponymy/hyperonymy* and *causes/is caused* relations, which are typically used for nouns and verbs.

BTB-WN was created in several steps. The work on it started on the base of the Core WordNet subset⁵ of Princeton WordNet (PWN) (Fellbaum, 1998) which contains the 5000 most frequent English senses. Later it was expanded with content words from the BulTreeBank and a Bulgarian frequency list, and also with encyclopedic data: senses from the Bulgarian versions of Wikipedia and Wiktionary. BTB-WN was successively expanded by several initiatives, such as expansion with multi-word expressions, named entities, terms from certain domains, etc. Originally BTB-WN was mapped to the PWN, but since 2020 the mapping is transferred to the Open English WordNet (EOW) (McCrae et al., 2020), because it is being currently maintained and developed, unlike PWN. Recently several new synset relations were introduced in BTB-WN in addition to the relations from PWN and OEW, which are also used with some modifications. The latest version of BTB-WN is 4.0⁶. It contains more than 33 000 synsets and is available for online browsing⁷. Soon BTB-WN 4.0 will be freely available for downloading in WordNet-LMF XML format.

Section 2 gives an overview of relevant studies about prepositions, Section 3 presents the semantic preposition classification that is used, Section 4 contains the synset model for prepositions in BTB-WN, and Section 5 introduces the plans for future work and Section 6 concludes the paper.

2 Related Works

The processing of prepositions is a challenging task for NLP mainly because of the high polysemy of prepositions. Yet there are many studies which attempt at sense disambiguation and semantic classification of prepositions.

Typically wordnets do not contain closed-class words including prepositions (currently, the only wordnet with prepositions is the BulNet for Bulgarian but they do not have relations and hierarchy⁸), but there are recent studies which show such attempts. Amaro (2018) presents an approach for incorporation of prepositions in wordnet, particularly of Portuguese prepositions for movement. The work provides tests for the establishment of several semantic relations between prepositions, following the relations for nouns, verbs, adjectives and adverbs in wordnets: *synonymy*, *antonymy*, *hyponymy*, *hypernymy*, *cause* and *is caused by*. These tests could be applied to prepositions in different languages. The approach towards prepositions in BTB-WN follows that of Amaro (2018), but with the aim to identify even more relations.

The research of Da Costa and Bond (2016) provides another proof for the benefits of enriching wordnets with different parts of speech by introducing non-referential concepts. They incorporate interjections, numeral classifiers and exclamatory pronouns in the Open Multilingual WordNet⁹ and establish several kinds of relations between them and the other parts of speech.

A resource fully dedicated to prepositions is the PrepNet (Saint-Dizier, 2008). It introduces a categorisation based on thematic roles particularly for French prepositions, but it is applicable also for English, Spanish and German. PrepNet distinguishes prepositions in two levels – abstract notion level and language realization level. The first one is conceptual and does not differ in languages and the second one regards the realizations in different languages.

The Preposition Project is a resource “designed to provide a comprehensive characterization of preposition senses suitable for use in natural language processing” (Litkowski and Hargraves, 2005). It uses

⁵<http://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt>

⁶<https://clada-bg.eu/bg/centers-and-services/language-technologies/btb-wordnet.html>

⁷<https://concordance.webclark.org/>

⁸<http://dcl.bas.bg/en/resursi/wordnet/>

⁹<https://omwn.org/>

the presentations of prepositions from three different sources and compares them – the Oxford dictionary of English, tagged prepositions from FrameNet and the prepositions from an English grammar. This way every preposition sense in the Preposition Project is described by a semantic role name, syntactic and semantic properties of its complement and attachment point, link to its definition in the dictionary, syntactic function and meaning from the grammar, different prepositions with a similar semantic role, frames and frame elements from FrameNet, other syntactic forms in which the semantic role may be found and the position of the given preposition in a network of prepositions.

Lassen (2006) presents a novel approach towards an ontology-based preposition processing – 16 preposition senses are used in the study to mark the semantic relations between noun phrases and the result of that is aimed at determining conceptual distance for information retrieval purposes.

The resource of Schneider et al. (2015) based on preposition supersenses provides a general purpose classification of prepositions, aimed at automatic WSD, where every supersense has detailed explanations, dictionary senses, example sentences and mapping to other resources, all of that for the benefit of the annotators.

O’Hara and Wiebe (2003) perform a WSD of prepositions based on the semantic role annotations of the Penn TreeBank and FrameNet. They use the semantic roles as senses of the prepositions. Moreover, they enrich the standard approach of WSD for using collocations by taking into consideration also WordNet hypernyms. High-level synsets are serving as collocations in the form of semantic categories. The two resources that are used are providing data with different levels of granularity – the Penn Treebank semantic roles classification is more compact (7 types) and that of FrameNet is very fine-grained (140 classes) (the classifications are presented in Section 3). The results show that the accuracy is higher with the less fine-grained roles from PennTreebank, but in the processing of the both resources the wordnet hypernyms prove beneficial. In the later research of O’Hara and Wiebe (2009) the topic of preposition disambiguation is further explored with the addition of data from more semantic role resources.

For the aims of predicting semantic relations Srikumar and Roth (2013) use the annotated prepositions from the SemEval 2007 shared task of WSD for prepositions and semantic roles from FrameNet. Their approach includes identifying the semantic types of the prepositions’ governor and object and for that wordnet hypernyms are used. The method is very beneficial, because for polysemous prepositions, the sense prediction is only possible with information for the arguments.

Similarly, Bailey et al. (2015) exploit VerbNet frames, wordnet relations and selectional restrictions with the purpose of resolving prepositional phrase attachment.

3 Semantic Classification of Prepositions

The semantics of Bulgarian prepositions are very well explored and many classifications are available. Several of them (Stoyanov, 1983; Konstantinova, 1982; Boyadzhiev et al., 1998) were consulted for the goal to adapt a more compact general-purpose classification of prepositions for the integration in the BTB-WN. The first categorisation (Stoyanov, 1983) is the most thorough – some prepositions there have around 30 senses, because the classification separates all subsenses and provides examples for rare, archaic, dialectal, etc. usages. The classification of Boyadzhiev et al. (1998) is on the contrary very generalized – it has 13 categories and does not include examples from different speech registers. The overview of the history of classifications of Bulgarian prepositions made by Konstantinova (1982) is also taken into consideration.

The resulting adapted classification contains 15 categories of prepositions: location, time, transition, manner and instrument of action, possession, quantity, degree and exceeding of limit, purpose, origin and part of a whole, opposition, comparison, cause and object class: exchange, exclusion, opinion and thought. Some modifications differing from the above-mentioned classifications are done in order to make the groups of prepositions in BTB-WN more compact.

For example, the closely related classes `manner of action` (вЪРВЯТ в редица *vървят v reditsa* “they walk in a line”) and `instrument of action` (ЯМ с вилица *jam s vilitsa* “I eat with a fork”) here are united in one category.

The same approach is applied to the *origin* (изследовател от България *izsledovatel ot Bălgarija* “researcher from Bulgaria”) and *part of a whole* (филм в две части *film v dve časti* “movie in two parts”) classes.

The approximation of time (около 9 часа *okolo 9 časa* “around 9 o’clock”) and approximation of quantity (към 3 километра *kăm 3 kilometra* “about 3 kilometers”) classes from Stoyanov (1983) here are generalised and included respectively in the *time* and *quantity* classes.

The exceeding of limit sense (това е свръх силите ни *tova e svrăh silite ni* “this is beyond our powers”) is considered a part of the *quantity* category.

An *object superclass* is created to unite the expression of relations for *exchange* (ще го направя вместо теб *šte go napravja namesto teb* “I will do it instead of you”), *exclusion* (яке без копчета *jake bez kopčeta* “jacket without buttons”), *thought* (разсъждавам върху проблема *razsăždavam vărhu problema* “I reflect on the problem”) and *opinion* (за мен това е най-доброто решение *za men tova e naj-dobroto rešenje* “for me this is the best decision”). The *prep.obj.thought* class includes expression of object of thought, speech and writing.

Additionally, a decision is made for the metaphorical usages of a given class to be considered part of it, not a separate group. For instance, usages like *тя го чу сред всички гласове* *tja go ču sred vsichki glasove* “she heard him among all the voices” are considered as examples of the *location* class.

Table 1 shows the semantic classes and synset categories of the prepositions and the preposition distribution in them.

Semantic Class	Synset category	Prepositions
locative	prep.location	в (във), връз, всред, въз, върху, до, за, зад, из, извън, иззад, измежду, изпод, край, към, между, на, над, накрай, наред, низ, о, около, от, отвъд, откъм, отсам, оттам, отгатак, по, под, подир, подире, покрай, помежду, посред, пред, през, при, против, пряко, след, спроти, сред, срещу, у
temporal	prep.time	в (във), всред, до, за, край, към, между, на, накрай, наред, около, от, по, подир, подире, покрай, помежду, посред, пред, преди, през, при, с (със) след, спроти, сред, срещу, у
manner and instrument of action	prep.manner	без, в (във), като, на, по, под, посредством, с (със), според, чрез
cause	prep.cause	за, заради, от, оттам, по, поради, пред
purpose	prep.purpose	до, за, заради, към, поради
possession	prep.possession	на, от, с (със), у
origin and part of a whole	prep.origin	в (във), на, от
quantitative, degree and exceeding of a limit	prep.quantity	до, за, към, между, на, над, около, от, по, под, с (със), около, свръх
exchange	prep.obj.exchange	вместо, за, заради, наместо, срещу, спроти
exclusion	prep.obj.exclusion	без, извън, освен
opinion	prep.obj.opinion	за, според
thought	prep.obj.thought	върху, връз, въз, до, за, заради, към, над, около, по, спрямо
transition	prep.transition	в (във), на, от
comparison	prep.comparison	като
opposition	prep.opposition	въпреки, против, пряко, спроти, срещу

Table 1: Semantic classes and synset categories of prepositions.

3.1 Parallel with Classifications for Other Languages

A comparison of the semantic classification of Bulgarian prepositions could be made with the semantic roles used in the Penn Treebank for the prepositional phrases. They determine the semantic relation which the prepositions express. In Penn Treebank there are seven types of semantic roles for prepositional phrases – beneficiary, direction, spatial extent, manner, location,

purpose/reason and temporal. Most of them correspond to the classes of the Bulgarian classification, but with several differences: there the purpose and reason classes are united in contrast with the Bulgarian; also the location relation here is divided in more fine-grained subtypes – spatial extent and location; and the direction relation here is a separate class, but in the Bulgarian categorisation it could be either in the purpose or location classes, depending on the given instance.

As mentioned above, the thematic roles-based classification of PrepNet (Saint-Dizier, 2008) is applicable for several languages, and additionally it is observed that it has many similarities with the Bulgarian classification.

It contains the following senses: localization (with subsenses source, destination, via/passage, fixed position, quantity (numerical or referential quantity, frequency and iterativity, proportion or ratio), manner (manners and attitudes, means (instrument or abstract), imitation or analogy), accompaniment (adjunction, simultaneity of events, inclusion, exclusion), choice and exchange (exchange, choice or alternative, substitution), causality (cause, goal or sequence, intention), opposition, ordering (priority, subordination, hierarchy, ranking, degree of importance) and minor groups like about, in spite of and comparison.

The generalized classifications with small number of classes, such as that of Penn Treebank, surely have benefits, as it is proved by the experiments of O'Hara and Wiebe (2003) for WSD of prepositions on the base of the semantic roles from Penn Treebank and FrameNet. The classification of FrameNet is much more fine-grained with more than 140 types of semantic roles and the results of the research show that the accuracy of the WSD is higher with the data from Penn Treebank. However, the semantic roles of FrameNet are frequently used for preposition disambiguation and processing (see Section 2).

In O'Hara and Wiebe (2003) the top 25 roles from FrameNet are sorted: speaker, message, self-mover, theme, agent, goal, path, cognizer, manner, source, content, experiencer, evaluatee, judge, topic, undefined, cause, addressee, perceptual source, phenomenon, reason, area, degree, body part, protagonist. Even though in FrameNet the roles distinguish very finely the different subsenses, similar types with the Bulgarian classification could be observed: in both of them there are manner and source (in the Bulgarian it is formulated as origin class) types; the cause class in FrameNet is divided in two – cause and reason; goal is corresponding to the purpose class; degree is included in the quantity class in the Bulgarian categorisation; area in FrameNet is a subtype of the more general location class, etc.

The classification used by Lassen (2006) also has similarities with the Bulgarian one. Some of the similar classes (concerning the same senses but rather different in formulation) that they share are temporal aspects, location, position, purpose, function, cum (for accompaniment, etc.), causes, caused by, by means of, instrument, via, comprising, has part and part of. There are also several classes regarding acts and processes (agent/patient/source/result/destination of act or process), which are not available in the Bulgarian classification.

The work of Srikumar and Roth (2013) presents an inventory of 32 preposition relations, many of which match with Bulgarian classes, such as cause, location, manner, purpose, etc. However, the relations are generally more fine-grained than in our classification, for example destination and direction classes are separate, they are not considered as a part of a broader location category. Based on the preposition relations of Srikumar and Roth (2013) and the thematic roles from VerbNet, Schneider et al. (2015) introduce a classification of 73 fine-grained preposition supersenses, which form a hierarchical taxonomy.

4 Preposition Synset Model

The main intention towards the structure and relations of the preposition synsets is that they follow the model of all the other synsets in BTB-WN (for nouns, verbs, adjectives and adverbs) as much as possible,

given the differences between them.

Preposition synsets have synset category (which shows their semantic category), detailed definition, examples, synonyms if available and relations. In BTB-WN preposition synsets have the part of speech value *p*, following the format of the Global WordNet Association¹⁰ where this value is planned for adpositions.

Currently six type of relations between preposition synsets and between a preposition synset and other parts of speech are established – *synonymy*, *antonymy*, *hyponymy*, *hyponymy*, *similarity* and *semantic derivation*. The relations are mainly semantic, but also two derivational are used – *sem-derived-from-p* which links the preposition with the noun or adverbs that it is derived from and the opposite – *sem-derives-to-p*¹¹. Additionally, it is planned to introduce relations between verbs and prepositions and nouns and prepositions that link combinations of these two parts of speech to express a given meaning.

Figure 1 presents an example for the model with the synset of върху *vărhu* “over, on” (*prep.location*) which has several synonyms, the first two being archaic (врѣз, въз *vřáz, vřáz*); one *antonymy* relation with под *pod* “under”; it has *derivational* relations with the noun връх *vřáh* “top” and the adverb сврѣх *svřáh* “above, beyond”; it has a *similar* relation with the preposition над *nad* “over, above” that also expresses a spatial position higher than something.

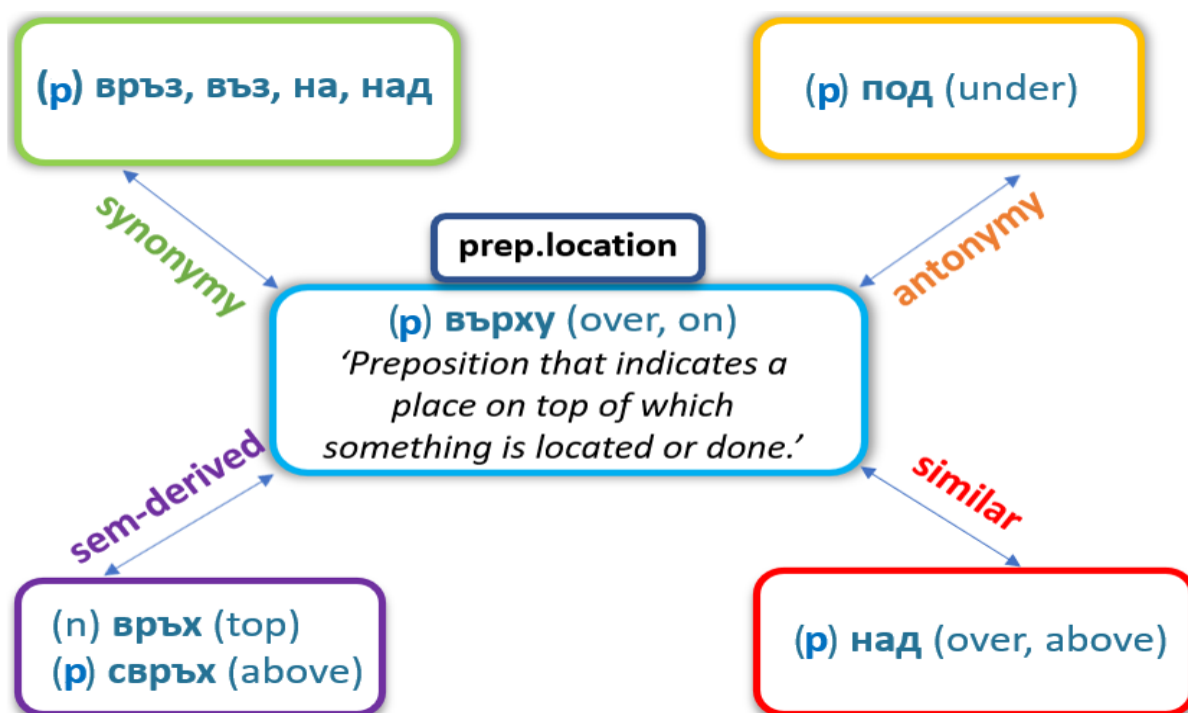


Figure 1: Example of the preposition synset model in BTB-WN

Synonymy is observed, for example, in the prepositions преди *predi* and пред *pred* “before” that are synonyms in their temporal sense and are united in one synset. The synset with the most synonyms – 7 – is from the manner class: в, на, по, под, с (*със*), посредством, чрез *v, na, po, pod, s (săs), posredstvom, črez* (“in”, “on”, “under”, “with”, “through”) meaning *expression of a way, form of doing, course of something*. (see Figure 2).

The *antonymy* relation links the synsets of без *bez* “without” (in the sense of *expression of lack from a part of a whole, prep.obj.exclusion*) and с *s* “with” (*expression of a belonging part or quality of something, someone, prep.possession*). An example for the *hyponymy* and *hyponymy* relations could be made with the synset for в, на, у *v, na, u* “in, at” (*expression of position in time, a moment when*

¹⁰<https://globalwordnet.github.io/schemas/>

¹¹More information about the newly introduced relations could be found in the BTB-WN Guidelines – <https://clada-bg.eu/en/centers-and-services/language-technologies/btb-wordnet.html>

Example	Lemma	ID	Order	Part of speech
Вървят @@@ в @@@ стройни редици.	в	217662	1	p
	на	217672	2	p
	по	217673	3	p
	под	217702	4	p
	с	217709	5	p
	посредством	217710	6	p
	чрез	217711	7	p
	със	217712	8	p

Figure 2: The preposition synset with the most members in BTB-WN.

something happens, prep.time), which is the hypernym of several narrower temporal senses such as от of “from” (*expression of initial limit in time, beginning of something*, prep.time), по, към, около по, към, около “by, around” (*expression of approximate time*, prep.time) and a few others, that are shown in Figure 3.

Example	Lemma	ID	Order	Part of speech
@@@ в @@@ 10 часа.	в	217662	1	p
	на	217672	2	p
	у	217663	3	p

Relations: **Hyponym chain** | Additional information | Tickets | Open ticket | Temporary notes

Изразяване на положение във времето, момент, отрязък или период от време, когато нещо се случва, извършва.

hyponym

Lemma	Definition	Part of speech	EquiID
сред,сред,насред,поср	Изразяване на момент в средата на определено време.	p	155916
по,към,около	Изразяване на приблизителност по време, приблизителен момент, в който става нещо.	p	155921
накрай	Изразяване на край на някакъв момент, период.	p	155922
от	Изразяване на начален предел във времето, започване на нещо.	p	155923
през	Изразяване на период от време, в границите на който се случва, извършва нещо.	p	155927
при	Изразяване на положение във времето спрямо други придружаващи събития, обстоятелства, на и	p	155929
срещу,спроти	Изразяване на извършване, осъществяване на нещо в навечерието на нещо друго, близост по вре	p	155931

Figure 3: Example of *hyponymy* and *hyponymy* relations in the CLaDA-BG Dict – the editing system for BTB-WN

Figure 4 shows a locative sense of в в “in” (*expression of location, place where something is, happens or is done*, prep.location). This is the preposition synset with most numerous hyponyms so far – 10.

The *similarity*¹² relation is established for example between до, пред, при do, pred, pri “to, in front of,

¹²Similarity relation is used in PWN to link closely related senses, originally only adjectives – <https://globalwordnet.github.io/gwadoc/#similar>

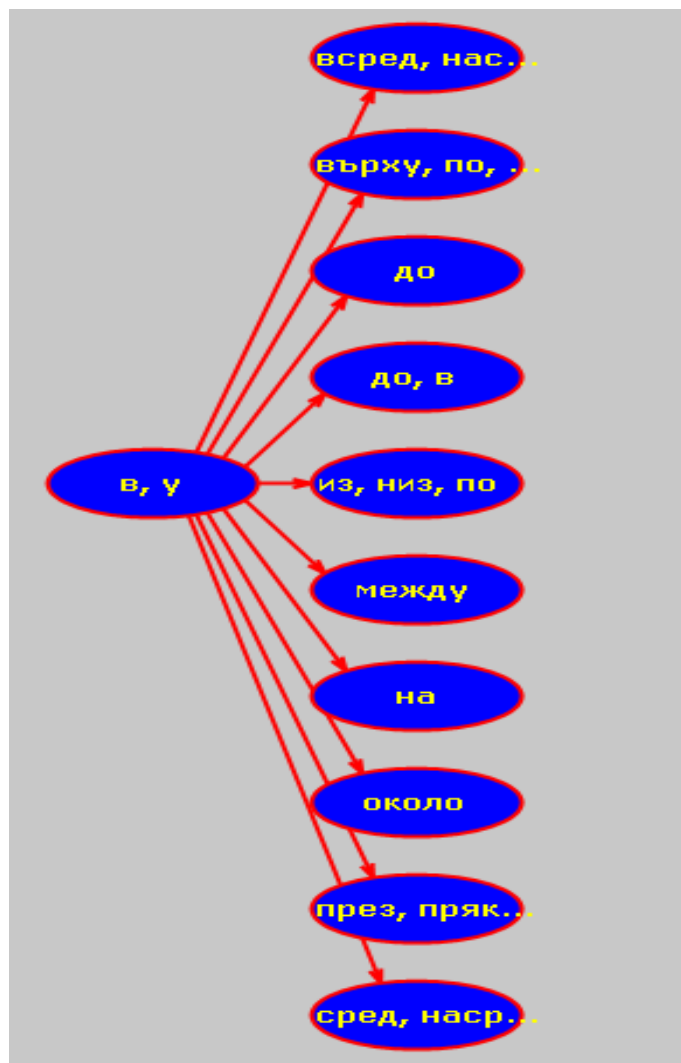


Figure 4: The preposition synset with the most hyponyms in BTB-WN – в “in”

at” (*expression of close proximity with something or someone in, during movement, prep.locative*) and зад, след *zad, sled* “behind, after” (*expression of close proximity with something or someone, prep.locative*).

The *sem-derived-from-p* and *sem-derives-to-p* are used for cases such as след *sled* “after, behind” (*prep.locative* and *prep.time*) that are derived from the noun следа *sleda* “trace” (*a visible mark left by the passage of person, animal or vehicle*) (see Figure 5).

In terms of origin Bulgarian prepositions are traditionally classified in two groups: simple and compound. The compound prepositions can be derived from nouns (typically they are related with a certain case form of the noun) or from adverbs. Derivational relations which link adjectives and nouns, adverbs and verbs and, adverbs and adjectives (*sem-derived-to/from*) were recently introduced in BTB-WN 4.0 and are now extended also for linking prepositions with related nouns and adverbs.

62 preposition lemmas were integrated in BTB-WN, forming a total of 105 synsets. The highest polysemy is found in the prepositions на *na* (most frequently could be translated as “on”, “of”, “in”, etc.) with 12 synsets, followed by по *po* (“over”, “in”, “on”, etc.) with 11 synsets. Other prepositions with multiple senses are за *za* (“for”, “to”, “about”, etc.) and от *ot* (“from”) that are part of nine synsets each; с *s* (“with”) is in eight synsets and до *do* (“to”, “until”, etc.) and в *v* (“in”, “at”) are found in seven.

Lemma
10 examples, 7 of which to lemmas

Example	Lemma	ID	Order	Part of speech
	следа	12115	1	n
Потерите се придружаваха от хрътки, при	диря	12114	2	n
	отпечатък	12116	3	n

relations Hypernym chain Additional information Tickets Open ticket Temporary notes

Знак, оставен от преминаването на човек, животно, превозно средство и други.

sem-derives-to-p

Definition	Lemma	Part of speech
Изразяване на непосредствена близост при движение.	зад,след	p
Изразяване на посока на действие или движение, която следва посоката на нещо друго.	по,след,подир,подире	p
Изразяване на по-късна позиция във времето, случване, извършване на нещо по-късно от друго.	след,подир,подире	p

Figure 5: Example of *sem-derived-from-p* and *sem-derives-to-p* relations in the CLaDA-BG Dict – the editing system for BTB-WN

The category distribution of prepositions is shown in Table 1 and there it could be observed that the *locative* class has the most prepositions – 46 (which is not surprising, since this sense of the prepositions is considered to be their oldest and primary function), followed by *time* with 28 and *quantity* with 13 prepositions. The *comparison* class proves to be smallest – this sense is expressed only by one preposition: *като* *kato* “like”.

5 Future Work

The methods and experiments from related works on prepositions for NLP provide several beneficial directions for future work. Bulgarian prepositions are planned to be analysed on the base of the phrase types that they participate in. Data from a Bulgarian valency lexicon and the BulTreeBank will be derived. The research of Harabagiu (1996) could be considered for this task, because it provides approach for prepositional disambiguation with the use of information from wordnet (semantic relations, noun and verb categories, glosses) as well as the work of Anand Kumar et al. (2015) aimed at preposition disambiguation for machine translation, where hypernyms and lexicographer files (that contain information for POS, category, etc.) from the PWN are used.

The prepositional classification could be additionally validated with corpus analysis following the bottom-up approach for manual annotation of Lassen (2006) but instead of ontological types for the nouns in the phrases wordnet categories could be used, as in the approach of Srikumar and Roth (2013). This information for the different parts of speech in the prepositional phrases would be used for establishing more relations with the preposition synsets in BTB-WN.

Regarding the semantic relations in wordnet, more resources are planned to be consulted. To a greater extent than the grammars, explanatory dictionaries provide information for the dialectal, archaic and colloquial variants of prepositions. They also include information about synonymy and antonymy of prepositions, which could be used for establishing the corresponding relations in BTB-WN. This data is planned to be included in BTB-WN and thus to provide a more exhaustive presentation of prepositions. The different speech register variants of all the other parts of speech in BTB-WN are present, so the approach towards the prepositions would be correlating. Etymological dictionaries would be used for establishing more *sem-derived-from/to* relations with the words that prepositions are derived from.

Additionally, the more complex in structure types of prepositions are still not taken into consideration in this research, so a next task could be the processing of polyprepositional constructions and also of prepositions with grammatical functions could be explored.

6 Conclusion

The paper presents the initial stages of the attempt at integrating closed-class words, namely prepositions, in the BTB-WN. Wordnets usually do not include prepositions in their structure, however relevant studies provide evidence that this task is possible and beneficial.

So far for BTB-WN are developed a semantic classification of prepositions and a synset model with several semantic relations. The paper outlines the directions for analysis and further processing of prepositions. The goal of this attempt is to improve the application of BTB-WN for semantic annotation and to use it as a resource for creating Bulgarian language models. Since prepositions express relational information, they have a key role for the semantic interpretation.

Acknowledgements

This work was supported by *the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG*, Grant number DO01-377/18.12.2020. and the Grant No. BG05M2OP001-1.001-0003, financed by *the Science and Education for Smart Growth Operational Program (2014-2020)* and co-financed by *the European Union through the European structural and Investment funds*.

References

- Amaro, R. 2018. Integrating Prepositions in Wordnets: Relations, Glosses and Visual Description. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan.
- Anand Kumar, M., Rajendran, S. and Soman, K. P. 2015. Cross-Lingual Preposition Disambiguation for Machine Translation. *Procedia Computer Science*, 54:291-300.
- Bailey, D., Lierler, Y. and Susman, B.. 2015. Prepositional Phrase Attachment Problem Revisited: how Verbnet can Help. *Proceedings of the 11th International Conference on Computational Semantics*, ACL, London, UK.
- Baldwin, T., Kordoni, V. and Villavicencio, A. 2009. Prepositions in Applications: A Survey and Introduction to the Special Issue. *Computational Linguistics*, 35(2):119-149. MIT Press, Cambridge, MA.
- Bond, F. and Da Costa, L. M. 2016. Wow! What a Useful Extension! Introducing Non-Referential Concepts to Wordnet *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), Portorož, Slovenia.
- Boyadzhiev, T., Kutsarov, I., and Penchev, Y. 1998. Contemporary Bulgarian language. Phonetics, lexicology, word formation, morphology, syntax. Petar Beron, Sofia, Bulgaria.
- Fellbaum, C. 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.
- Fillmore, J. C., Wooters, C. and Baker, C. F. 2001. Building a Large Lexical Databank Which Provides Deep Semantics. *Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation*, City University of Hong Kong, Hong Kong, China.
- Harabagiu, S. M. 1996. An application of WordNet to prepositional attachment. *Proceedings of the 34th annual meeting on Association for Computational Linguistics (ACL 1996)*. Association for Computational Linguistics, Santa Cruz, USA.
- Konstantinova, V. 1982. Prepositions in Bulgarian grammar literature. Publishing house of BAS, Sofia, Bulgaria.
- Lassen, T. 2006. An Ontology-based View on Prepositional Senses. *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*. ACL, Trento, Italy.

- Litkowski, K. and Hargraves, O. 2005. The Preposition Project. *ACL-SIGSEM Workshop on "The Linguistic Dimensions of Prepositions and Their Use in Computational Linguistic Formalisms and Applications"*. Colchester, England.
- Marcus, M., Kim, G., Marcinkiewicz, M., MacIntyre, R., Bies, A., Ferguson, M., Katz, K. and Schasberger, B. 1994. The Penn Treebank: Annotating predicate argument structure. *Proceedings of the ARPA Human Language Technology Workshop*, Plainsboro, NJ.
- McCrae, J. P., Rudnicka, E. and Bond, F. 2020. English WordNet: A new open-source WordNet for English. *K Lexical News*, 28:37-44.
- O'Hara, T. and Wiebe, J. 2009. Exploiting Semantic Role Resources for Preposition Disambiguation. *Computational Linguistics*, 35:151–184. MIT Press, Cambridge, MA.
- O'Hara, T. and Wiebe, J. 2003. Preposition Semantic Classification via Penn Treebank and FrameNet. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, ACL, Edmonton, Canada.
- Osenova, P. and Simov, K. 2018. The data-driven Bulgarian WordNet: BTBWN *Cognitive Studies – Études cognitives*, volume 2018(18). Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland.
- Prasad, R., Webber, B., Lee, A. and Joshi, A. 2019. The Penn Discourse Treebank 3.0 Annotation Manual. *Linguistic Data Consortium*, Philadelphia, USA.
- Saint-Dizier, P. 2008. Syntactic and Semantic Frames in PrepNet. *Proceedings of the Third International Joint Conference on Natural Language Processing*, volume II. Asian Federation of Natural Language Processing.
- Schneider, N., Srikumar, V., Hwang, J.D., and Palmer, M. 2015. A Hierarchy with, of, and for Preposition Supersenses. *Proceedings of the 9th Linguistic Annotation Workshop*, ACL, Denver, Colorado, USA.
- Srikumar, V. and Roth, D. 2013. Modeling Semantic Relations Expressed by Prepositions. *Transactions of the Association for Computational Linguistics*, 1:231–242, MIT Press, Cambridge, MA.
- Stoaynov, S. 1983. Grammar of contemporary Bulgarian standard language. Morphology. Publishing house of BAS, Sofia, Bulgaria.

Neural Metaphor Detection for Slovene

Matej Klemen Marko Robnik-Šikonja

University of Ljubljana, Faculty of Computer and Information Science

Večna pot 113, Ljubljana, Slovenia

{matej.klemen, marko.robnik}@fri.uni-lj.si

Abstract

Metaphors are linguistic expressions using comparison with another concept to potentially improve the language expressivity. Due to relevant downstream applications, metaphor detection is an active topic of research. Most of the research is focused on English, while other languages are less covered. In our work, we focus on Slovene, presenting the first word-level metaphor detection experiments. We apply multiple transformer-based large language models on four versions of two publicly available Slovene corpora: KOMET and G-KOMET. We perform monolingual, multilingual, and cross-lingual experiments, using the VU Amsterdam metaphor corpus as an additional source of metaphor knowledge. We evaluate the models quantitatively using word-level F_1 score and find that (1) the most consistently well-performed model is the trilingual CroSloEngual BERT model, (2) the addition of English data in multilingual experiments does not improve the performance significantly, and (3) the cross-lingual models achieve significantly worse results than their monolingual and multilingual counterparts.

1 Introduction

A metaphor is an expression that uses a comparison with another concept for rhetorical effect. For example, instead of saying “*his words offended me*” we might say “*his words cut deeper than a knife*”, comparing the effect of offensive words to the physical pain of a knife cut. Metaphors are ubiquitous in language and add color to conversations. The ability to detect and form metaphors can be applied to creative writing, such as news headline generation or rephrasing (Stowe et al., 2021), and enables the analysis of public discourse and evolution of language (Prabhakaran et al., 2021; Zwitter Vitez et al., 2022; Kutuzov et al., 2018).

Although humans are able to detect and understand metaphors with relative ease, metaphor detection has proven to be a challenging problem for computational models (Strapparava, 2018). Most existing work on metaphor detection has dealt with broadly-spoken languages such as English. The earlier approaches relied on handcrafted features such as abstractness and concreteness features in combination with machine learning models (Turney et al., 2011). Recent work is shifting towards the use of large language models that model the word context (Choi et al., 2021).

An example of a language with less researched metaphors is Slovene, where little work has been done on metaphor detection, although annotated datasets exist (Antloga, 2020; Antloga and Donaj, 2022). The only computational approach to metaphor detection for Slovene was done by Zwitter Vitez et al. (2022). Authors detect metaphors at the sentence-level by first training a model for idiom detection and then tuning it for metaphor detection on the KOMET corpus (Antloga, 2020). The idea behind their work is that although metaphors and idioms are different concepts, both are forms of figurative language that heavily rely on a word’s context.

In our work, we approach the metaphor detection at a fine-grained (i.e. word) level and present the results of initial experiments applying state-of-the-art language models to Slovene metaphors, leveraging multiple CLARIN resources in the process. Although metaphors are multi-word units, we model their

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

detection at a word level as the metaphors in the datasets are annotated at this level, following the MIPVU annotation scheme (Steen et al., 2010). In contrast to detection at the sentence level, this approach enables a clearer insight into what part of the text is metaphorical. On the other hand, such an approach runs the risk of the model only learning to detect the “easier” subset of words in a multi-word expression as the model only considers dependence between words implicitly, an aspect we discuss in our evaluation. We perform experiments on the KOMET metaphor dataset and the previously unexplored spoken language dataset G-KOMET. We test metaphor detection in a monolingual, multilingual, and cross-lingual setting, using the VU Amsterdam metaphor corpus (VUAMC) (Krennmayr and Steen, 2017) as an additional source of knowledge:

- In the monolingual experiments, we test the performance of large language models on word-level metaphor detection setting for Slovene.
- In the multilingual experiments, we test if the inclusion of English language can improve the performance on Slovene, for example by acting as a regularization mechanism.
- In the cross-lingual experiments, we test if the models are able to learn any transferable language-agnostic ideas behind metaphors on the English data and test it on the Slovene data.

We make the code to rerun our experiments publicly available¹.

The remainder of our paper is structured as follows. In Section 2, we describe the relevant existing work on metaphor detection. In Section 3, we describe the data used in our experiments. Section 4 presents our approach for metaphor detection, while the results are contained in Section 5. We present conclusions and discuss possible future directions in Section 6.

2 Related Work

The work on metaphor detection has progressed from initial manually engineered feature-based approaches in combination with machine learning algorithms towards increasingly more automated (end-to-end) feature learning in combination with deep neural networks.

In the initial work on metaphor detection, authors have tested various features such as concreteness of words, named entity features, and part of speech tag features (Tsvetkov et al., 2013; Beigman Klebanov et al., 2014) in combination with machine learning algorithms such as logistic regression (Cox, 1958). With the introduction of word embedding algorithms, authors have discovered that the embeddings are able to replace or simplify the process of feature construction. For example, Do Dinh and Gurevych (2016) use word2vec word embeddings (Mikolov et al., 2013) in their metaphor detection system and demonstrate their feasibility on English data. In addition to word embeddings, they also use part of speech tag embeddings, replacing the manual linguistic features with automatically learned ones. Although powerful, in contrast to the inherently contextually dependent nature of metaphors, earlier proposed word embeddings such as word2vec cannot produce a contextual word representation. To amend this, researchers have incorporated the embeddings in combination with models that are capable of capturing an increasing amount of context, such as long short term memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997) and convolutional neural networks (CNNs) (LeCun et al., 2010). For example, Pramanick et al. (2018) propose an approach using a bidirectional LSTM and conditional random fields (Lafferty et al., 2001) to improve the English metaphor detection performance. Lately, authors have started replacing LSTMs and CNNs with transformer models (Vaswani et al., 2017) due to their wide success on other tasks in natural language processing. Despite the powerful representation capability of transformer-based models, latest metaphor detection models continue to benefit from the inclusion of external information such as concreteness (Alnafesah et al., 2020) or the customization of architecture to account for the specifics of the metaphor detection task. For example, Choi et al. (2021) include custom components in their approach that model the process used in the annotation of metaphor datasets. They show that their system performs better than an out-of-the-box transformer model.

¹<https://github.com/matejklemen/metaphor-recognition>

Most of the work is focused on processing English metaphors, and the topic has been the focus of multiple shared tasks (Leong et al., 2018; Leong et al., 2020; Saakyan et al., 2022). Some other broadly spoken languages such as Chinese and Russian have also received significant attention and customized architectures (Song et al., 2021; Lu and Wang, 2017; Badryzlova et al., 2022), while work on other languages is relatively scarce in comparison. Examples of metaphor detection systems can be found for diverse languages, such as Spanish (Sanchez-Bayona and Agerri, 2022), Greek (Florou et al., 2018), and Uyghur (Qimeng et al., 2021), although they are typically focused on a narrower domain (e.g., literary) or present initial studies on the feasibility of metaphor detection.

With our work, we expand the research on non-English metaphor detection. The only previous approach for Slovene metaphor detection is by Zwitter Vitez et al. (2022), who detect metaphors at the sentence level using the KOMET dataset. In contrast, we approach metaphor detection at a more fine-grained (word) level and test the state-of-the-art transformer models in a monolingual, multilingual and cross-lingual setting.

3 Metaphor Datasets

In our experiments, we use three datasets for evaluating the metaphor detection performance: KOMET (Antloga, 2020), G-KOMET (Antloga and Donaj, 2022), and VUAMC (Krennmayr and Steen, 2017). In addition to these original datasets, we create modified versions of KOMET and G-KOMET containing “semantically interesting metaphors”, i.e. metaphors containing at least one noun or verb (marked as *NV* in Table 1).

The datasets are annotated using the MIPVU annotation scheme (Steen et al., 2010). Using the scheme, a word is annotated as “metaphor-related” if its contextual meaning differs from its basic meaning. In Table 1, we provide basic statistics of the three datasets.

	KOMET		G-KOMET		VUAMC
	full	NV	full	NV	
# documents	62		287		117
# sentences	13 963		5695		16 202
# words	259 881		52 955		238 509
# met.	(5.2%) 13 574	(1.2%) 3100	(1.1%) 560	(0.7%) 357	(9.5%) 22 620
# MRWi	13 191	2893	527	324	22 254
# MRWd	364	205	33	33	341
# MRWimp	19	2	0	0	25

Table 1: Statistics of the used datasets: number of documents, sentences, words, and metaphors, as well as types of metaphors (indirect - MRWi, direct - MRWd, implied - MRWimp). The columns marked NV show statistics for the processed version of datasets where only metaphors containing a noun or verb are kept as metaphors.

KOMET is a Slovene metaphor corpus containing 13 963 sentences. It contains journalistic, fiction and web texts extracted from the Slovene youth literature corpus MAKS (Verdonik et al., 2020). The corpus was annotated by one annotator. Approximately 5.2% of the words are marked as metaphors in the original version and approximately 1.2% in the NV version.

G-KOMET is a Slovene metaphor corpus containing 5695 sentences. In contrast to KOMET, it contains transcripts of spoken language extracted from the GOS corpus (Verdonik et al., 2013). The corpus was annotated by one annotator. Approximately 1.1% of the words are marked as metaphors in the original version and approximately 0.7% in the NV version.

VUAMC is an English metaphor corpus containing 16 202 sentences. It contains academic, news, conversational, and fiction texts from the BNC-Baby corpus (BNC Consortium, 2007). The corpus was annotated by four annotators.

In our experiments, we are interested in detecting the following three metaphor types annotated in all used corpora.

- Indirect metaphor (MRWi): a metaphor where a comparison is indirectly stated. For example, in “Tim se je zljajal nanj“ (“*Tim barked at him*”), Tim’s yelling is compared to the barking of a dog and Tim is indirectly compared to a dog.
- Direct metaphor (MRWd): a metaphor where a comparison is directly stated. For example, in “Tim je osel” (“*Tim is a donkey*”), Tim is being compared to a donkey to outline his stubbornness.
- Implied metaphor (MRWimp): a lexical unit that is not necessarily metaphorical by itself but refers to a previous metaphorically used word. For example, in “*Naturally, to embark on such a step is not necessarily to succeed immediately in realising it*”, “step” is related to a metaphor and “it” refers to “step”, so it is considered an implicit metaphor (Steen et al., 2010).

We use KOMET and G-KOMET for training and evaluation in mono-, multi-, and cross-lingual settings, while we use VUAMC for training in the multi- and cross-lingual setting. We do not evaluate the performance on English as there exist many works on the topic (Leong et al., 2020).

4 Metaphor Detection

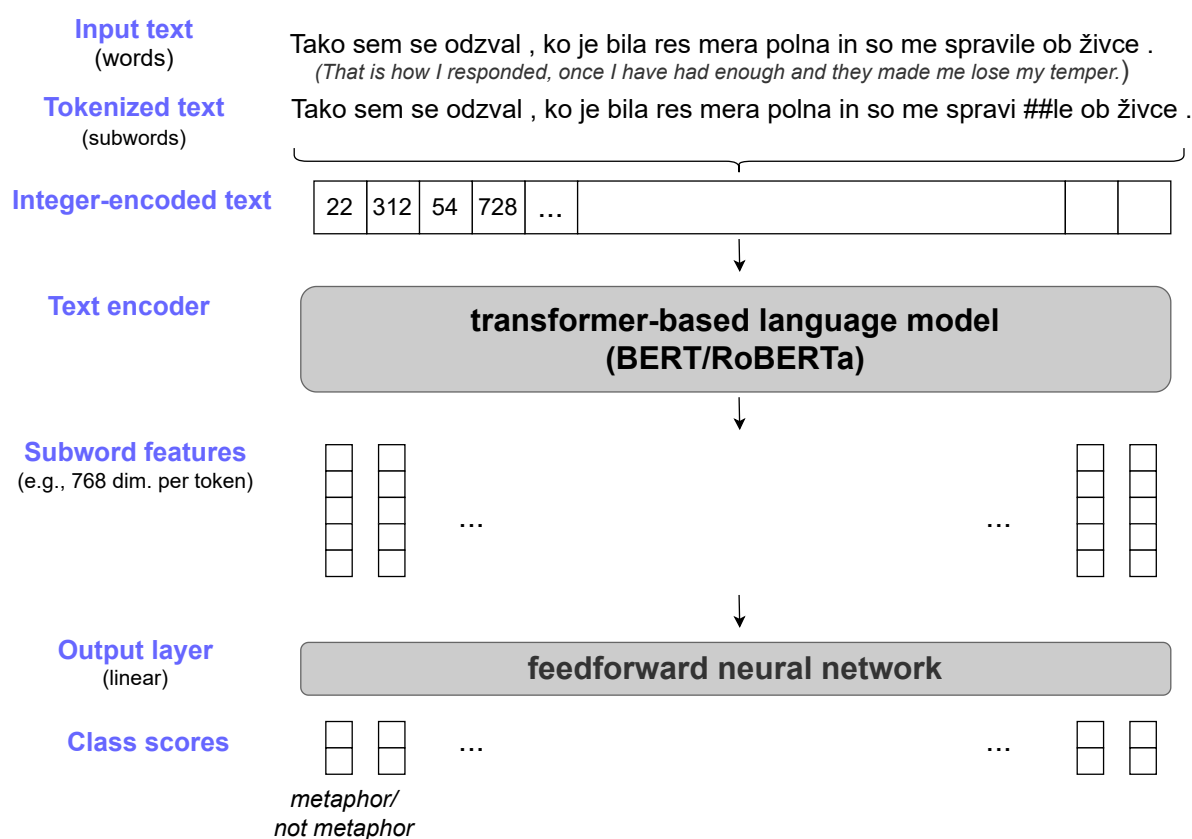


Figure 1: Schematic representation of the metaphor detection neural networks used in our experiments. Each integer-encoded subword of the input text is passed through the transformer-based language model to obtain a representation in the form of an embedding vector. Then, the embeddings are passed through the final classification layer to obtain a score of the subwords being a metaphor or non-metaphor.

We use the described datasets in a binary token classification setting, described next. The token classification procedure follows the best practice in existing literature (Devlin et al., 2019), to which we make

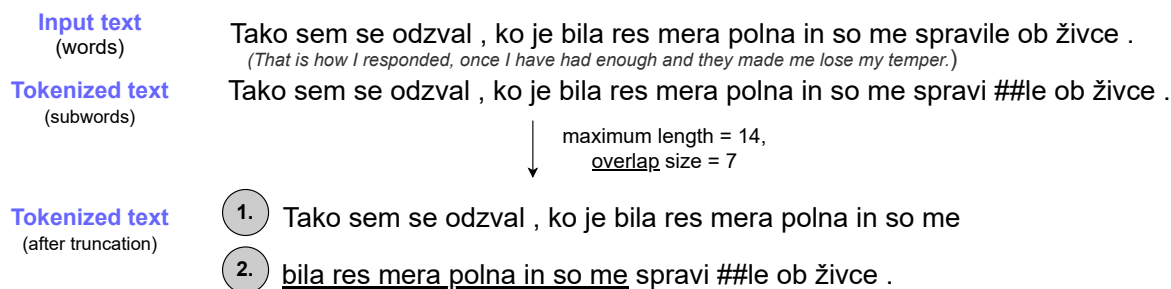


Figure 2: A toy example of the truncation logic in our modeling process. The input text gets tokenized into subwords and then broken up into two inputs as it is initially longer than the maximum model length (14 in this example). The second example has a partial overlap with the first example (marked with underline) of half the size of the maximum input length.

minor modifications to account for the nature of the metaphor detection task. In Figure 1, we show an example of detecting metaphors in a sentence.

The input text is first tokenized and encoded using the model-specific tokenizer, which converts the input into integer-encoded subwords. The subwords may be identical to the corresponding words or they can be smaller constituents of the words, depending on encoding of the input word in the model-specific vocabulary. The encoded text is then passed through the transformer-based model, which produces an internal representation in the form of a fixed dimensional (e.g., 768) vector for each subword. Intuitively, this representation captures the context in which a word appears. Throughout the training procedure, the representation is optimized to capture features that are useful for determining if a subword is a metaphor or not. In the final step, the token representations are passed through the feedforward neural network, which produces a score for each possible class, in our case metaphor and non-metaphor. By applying the softmax function on the scores, they can be interpreted as the probabilities of a subword being a metaphor or not. If the probability is higher than T , the subword is classified as a metaphor. We determine the probability threshold T on the validation set based on the F_1 score.

Transformer models have a limited maximum input length, i.e. the maximum number of subwords they accept as an input. The limit reflects primarily the computational constraints, more specifically the required amount of GPU memory and the maximal batch sizes used during model training. Inputs longer than the maximum length need to be handled separately. We break long input texts into multiple partially-overlapping input texts of the maximum allowed length. The size of the overlap is a hyperparameter and presents a trade-off: setting it too low drops potentially important context while setting it too high increases the amount of computation required for classification. We set it to half of the determined maximum length of a model as a sensible middle ground. In addition, we set the maximum input length high enough so that the examples do not need to be broken up frequently. The prediction for overlapping words is only taken into account once, i.e. when the overlapping segment is first observed. We show a toy example in Figure 2: the input “Tako sem se odzval, ko je bila res mera polna in so me spravile ob živce.” gets broken up into two inputs, with the second input partially overlapping the first one.

Another decision is how to classify words that the tokenizer splits into multiple subwords. For example, in Figure 1 the word “spravile” gets split into “spravi” and “le”. In the prediction phase, each subword is classified separately, so the predictions need to be aggregated to determine if a word is a metaphor. By default, we aggregate the predictions by setting the class of a word to the class of its first subword but further aggregation experiments are done in Section 5.4.

5 Evaluation

In this section, we analyze the performance of metaphor detection models on the described datasets. We describe the three evaluation settings in Section 5.1. Then we analyze the results quantitatively and

qualitatively in Sections 5.2 and 5.3. Last, we present an additional experiment analyzing the effect of subword score aggregation on the metaphor detection performance in Section 5.4.

5.1 Experimental Settings

In our experiments, we use four transformer models: monolingual SloBERTa (Ulčar and Robnik-Šikonja, 2021), trilingual CroSloEngual BERT (CSE BERT) (Ulčar and Robnik-Šikonja, 2020), and multilingual bert-base-multilingual-cased (mBERT_{BASE}) (Devlin et al., 2019) and XLM-RoBERTa_{BASE} (XLM-R_{BASE}) (Conneau et al., 2020). SloBERTa was trained on Slovene, CSE BERT on Slovene, Croatian, and English, and mBERT_{BASE} and XLM-R_{BASE} were trained on 104 and 100 languages, respectively. In terms of the number of parameters, SloBERTa and CSE BERT are comparable at 110 million parameters, mBERT_{BASE} is larger with 172 million parameters, and XLM-R_{BASE} is the largest with 270 million parameters. As the representation of the tokens (token features), we take the hidden state of the last transformer layer. In our preliminary experiments, we have also experimented with using a learned combination of all the hidden states, but found no significant difference on the metaphor detection performance. We first test all models in a monolingual setting, i.e. training them on the Slovene training set and evaluating them on the Slovene test set. We test the multilingual models also in a multi-lingual and cross-lingual settings, i.e. training them on the combined Slovene and English training set or just the English training set, and evaluating them on the Slovene test set. As the models use different tokenization which could lead to incomparable subword-level metrics, we use the word-level F_1 score of the positive class (i.e. *metaphor*) for evaluation.

We train the models for 10 (on KOMET and VUAMC) or 20 epochs² (on G-KOMET) and select the best model based on the validation set F_1 score. We use the learning rate $2e - 5$ and set the batch size to the maximum number that is possible on a 11GB GPU. We set the maximum input length of the models to the 99th percentile of the input lengths in the training set. This is different for every tokenizer-dataset pair but is typically between 80 and 100 subwords.

As we noticed a high variation in the results in our preliminary experiments, we perform the evaluation using 5-fold cross validation. The examples are split into folds on the document-level to reduce the possibility of an information leak. The folds are determined in a way that the proportion of metaphors is approximately the same in each fold. This simplification is due to the focus of the paper being the feasibility of metaphor detection on Slovene and not the ability of models to handle distribution shift. In each of the five evaluation runs, we set aside 10% of the training documents as the validation set. We split VUAMC in the same distribution-preserving fashion in the ratio 80%:20% training:test documents, although we do not evaluate models on English.

When comparing scores of different variants of a model, we use the Wilcoxon signed-rank test (Wilcoxon et al., 1970). We test the null hypothesis that the mean scores are the same using the significance level $\alpha = 0.01$.

5.2 Metaphor Word-Level Detection Results

In Table 2, we show the mean F_1 scores and standard deviations of the models. In one setting (SloBERTa on the G-KOMET_{NV} dataset), the model did not converge, so we mark its performance as *N/A* and exclude it from further analysis. In addition to the models mentioned in the previous section, we include a naïve baseline which predicts the metaphor class with probability 0.5 to check if the models have learned anything beyond random guessing. We see that the models in all cases surpass this baseline.

In all three settings (monolingual, multilingual, and cross-lingual), the F_1 scores on the original (i.e. full sized) datasets are higher than on the NV counterparts. For example, the highest mean F_1 score on KOMET_{full} is 0.607, while it is only 0.401 on KOMET_{NV}. This implies that a noticeable portion of the score on the full version comes from correctly detecting metaphors that are not nouns or verbs, e.g., adpositions such as “*na*” (“*on*” in English). These make up a significant amount of all the annotated metaphors, but are less interesting for practical applications such as rephrasing sentences with metaphors. The differences between full and NV variants on KOMET are higher than on G-KOMET as

²In preliminary experiments, we have found that the validation F_1 score kept increasing after 10 epochs, so we increased it.

Model	KOMET _{full}	KOMET _{NV}	G-KOMET _{full}	G-KOMET _{NV}
0.5/0.5 baseline	0.095 (0.008)	0.024 (0.003)	0.022 (0.003)	0.013 (0.001)
<i>(monolingual)</i>				
CSE BERT	0.606 (0.069)	0.361 (0.031)	0.261 (0.021)	0.243 (0.040)
SloBERTa	0.596 (0.071)	0.401 (0.040)	0.243 (0.021)	N/A
XLM-R _{BASE}	0.591 (0.073)	0.348 (0.027)	0.245 (0.033)	0.205 (0.049)
mBERT _{BASE}	0.575 (0.058)	0.304 (0.028)	0.216 (0.033)	0.173 (0.028)
<i>(multilingual_{EN+SL})</i>				
CSE BERT	0.607 (0.068)	0.389 (0.031)	0.313 (0.039)	0.276 (0.038)
XLM-R _{BASE}	0.599 (0.077)	0.354 (0.028)	0.282 (0.024)	0.283 (0.045)
mBERT _{BASE}	0.573 (0.065)	0.320 (0.020)	0.223 (0.016)	0.237 (0.039)
<i>(cross-lingual_{EN⇒SL})</i>				
CSE BERT	0.351 (0.035)	0.178 (0.036)	0.124 (0.041)	0.073 (0.021)
XLM-R _{BASE}	0.408 (0.046)	0.134 (0.021)	0.110 (0.010)	0.070 (0.010)
mBERT _{BASE}	0.374 (0.041)	0.112 (0.012)	0.089 (0.009)	0.053 (0.008)

Table 2: Mean word-level F_1 scores of metaphor detection models measured using 5-fold cross validation. The corresponding standard deviations are shown in parentheses. *N/A* indicates that the model did not converge, i.e. the validation metric did not improve in the training process. We mark the highest mean F_1 score in bold.

the proportion of metaphors is lower in the latter, so the proportion of metaphors in the NV version is not reduced as significantly as in KOMET. We further analyze the performance of models concerning part of speech tags of metaphors in Section 5.3.

In the monolingual experiments, we see that the models achieve comparable results with minor differences. In general, the best performing model across the four datasets is CSE BERT, achieving the best mean F_1 score on three (KOMET_{full}, G-KOMET_{full}, and G-KOMET_{NV}) and the second best mean F_1 score on one (KOMET_{NV}) dataset. However, the differences in performance between the models are not statistically significant.

In the multilingual experiments, the inclusion of English training data does not have a significant effect on the model performance. Comparing the multilingual models to their monolingual counterparts shows that there are potential differences in the performance on the G-KOMET dataset, although their statistical significance cannot be confirmed due to the high deviation in model performance. The improvement is possibly a consequence of the heavily imbalanced dataset and a smaller amount of training examples: G-KOMET is five times smaller than KOMET and contains only approximately 1% of metaphors.

In the cross-lingual experiments, we see a significant drop in performance across all the datasets. The results indicate that there is a limited amount of metaphor knowledge that can be transferred from the used English source to the Slovene target datasets. This is especially clear on G-KOMET, where an additional obstacle is the specialized domain, i.e. spoken language. Due to this, the worst performing cross-lingual models closely approach the random baseline performance, although the improvement in performance over the random baseline is still statistically significant.

5.3 POS-Tag Analysis of the Detected Metaphors

To observe what the models successfully detect, we analyze the model predictions based on universal part of speech (UPOS) tags of metaphors. We obtain them using the Trankit (Nguyen et al., 2021) library, using the Slovenian-SSJ large model. In Figure 3, we show the proportion of correctly detected metaphors

for each UPOS tag for the CSE BERT model on the four datasets in the monolingual setting. We select this model as it consistently performs well, but we have noticed that the proportions are similar for other models in the monolingual and multilingual setting. On $\text{KOMET}_{\text{full}}$, we observe that the model most accurately detects the adpositions (82.1% of them), which present the majority of the annotated metaphors. On the other hand, only 35% of nouns and 38.4% of verbs are correctly detected. The proportion of correctly detected noun and verb metaphors increases in the more narrowly focused KOMET_{NV} dataset. A similar conclusion can be drawn for G-KOMET, shown on Figure 3c and 3d, although adpositional metaphors do not present the majority in $\text{G-KOMET}_{\text{full}}$; instead, nouns and verbs present the majority. Therefore, the increase in the detection performance for nouns and verbs is significantly less noticeable in the $\text{G-KOMET}_{\text{NV}}$ dataset.

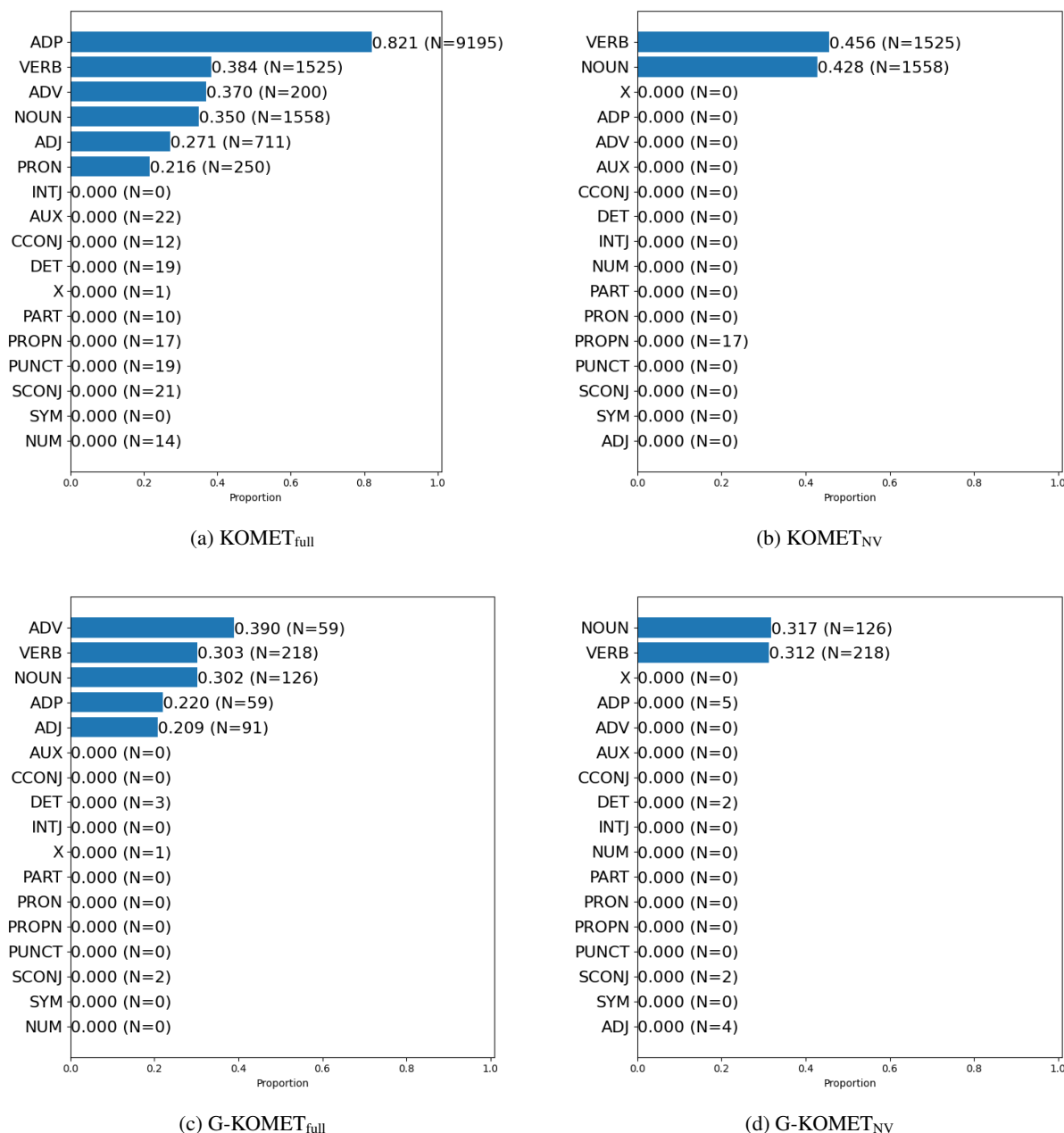


Figure 3: Proportion of metaphors correctly detected by CSE BERT in a monolingual setting, grouped by their UPOS tag.

We perform the same analysis for the best performing model in the cross-lingual setting on the KOMET dataset. The results are shown on Figure 4. Interestingly, although it performs significantly

worse in terms of F_1 score, this is primarily a consequence of poorly detecting the adpositional metaphors. On the other hand, the proportion of correctly detected nouns and verbs is equal or better than on KOMET_{NV} .

While it does improve the ability of detecting semantically interesting metaphors, ignoring non-noun and non-verbal metaphors is not the ideal solution as these may still be used in some linguistic analyses. More importantly, they can act as constituents of a metaphor composed of multiple words. For example, in the metaphor “spravile ob živce” (“to make someone lose their temper”), the word “ob” is essential for the phrase to be considered a metaphor. We note that the annotation of multi-word metaphors in existing datasets is imperfect as they are commonly annotated as multiple single-word metaphors: e.g., “spravile”, “ob”, and “živce” are labelled as three metaphors.

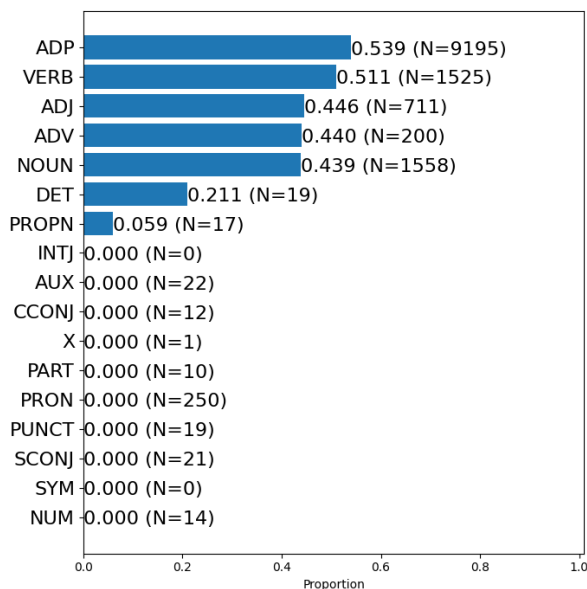


Figure 4: Proportion of metaphors correctly detected by $\text{XLM-R}_{\text{BASE}}$ on $\text{KOMET}_{\text{full}}$ in a cross-lingual setting, grouped by their UPOS tag.

5.4 Analysis of Subword Score Aggregation

In this additional experiment we test the effect of the strategy used to aggregate subword predictions. As the model tokenizers may break up the input word into multiple subwords, the predictions need to be aggregated to the word level. In our implementation, we use by default the prediction for the first subword as the prediction of the word - we refer to this strategy as *first*. However, as the subword predictions are made independently, meaning the prediction \hat{y}_i is independent of \hat{y}_{i-1} , it is possible that subwords belonging to the same word obtain inconsistent predictions. Therefore, we test two additional aggregation strategies to check if using a different strategy leads to a significant difference in performance:

- *majority*: the prediction of a word is determined as the majority prediction of its subwords;
- *any*: the prediction of a word is metaphor if any of the subwords is a metaphor, and non-metaphor otherwise.

To avoid overfitting the test set, we compare the validation set F_1 scores. We test this only for the $\text{XLM-R}_{\text{BASE}}$ model as its tokenizer is designed to handle 100 languages, so the words are far more likely to get divided into subwords than in the monolingual SloBERTa or trilingual CSE BERT model. From the results shown in Table 3, we can observe that the difference in performance across the three strategies is minimal and statistically insignificant. The results indicate that the model is likely to track the dependence between predictions in its hidden layers. As the results on $\text{XLM-R}_{\text{BASE}}$ show no difference,

we skip comparisons for other models as the corresponding tokenizers are equally or less likely to split words into subwords, so the results are unlikely to be different.

Model	Strategy	KOMET _{full}	KOMET _{NV}	G-KOMET _{full}	G-KOMET _{NV}
XLM-R _{BASE}	first	0.662 (0.035)	0.419 (0.024)	0.270 (0.021)	0.226 (0.045)
	majority	0.663 (0.036)	0.418 (0.021)	0.266 (0.009)	0.219 (0.049)
	any	0.664 (0.037)	0.420 (0.032)	0.272 (0.026)	0.227 (0.056)

Table 3: Comparison of subword prediction aggregation strategies. The prediction for a token is either the prediction for its first subword (*first*), the majority prediction of all its subwords (*majority*), or determined as the metaphor if at least one of the subwords is a metaphor, and non-metaphor otherwise (*any*). The scores are mean word-level F_1 scores measured using 5-fold cross validation. The corresponding standard deviations are shown in parentheses.

6 Conclusion

We presented the results of the first word-level metaphor detection attempt on Slovene data, analyzing the performance of the models in a monolingual, multilingual, and cross-lingual setting. Our approach considers metaphor detection as a standard token classification task with minor modifications, such as prediction aggregation and threshold optimization, to account for the specifics of the task. The results show that the models have plenty of room for improvement and perform best at detecting semantically less interesting metaphors, such as adpositions. The inclusion of English data in multilingual experiments has a minor and insignificant effect. The performance drops significantly in the cross-lingual experiments, indicating that there is a limited amount of knowledge that is transferrable from English to the Slovene datasets.

An issue in current metaphor datasets is the disjoint nature of metaphor annotations, i.e. multi-word metaphors are commonly annotated as multiple single-word metaphors which potentially limits options to improve their modeling. In future work, we plan to tackle this issue by proposing an automatic grouping mechanism which will allow modeling metaphor detection as a span extraction task instead of a token classification task.

An additional direction for future work is the introduction of new datasets as resources large enough to train neural networks are only available for a limited set of languages. The introduction of new datasets will enable detection in new domains and languages as well as potentially enable cross-lingual transfer.

Acknowledgements

The research was supported by the Slovene Research Agency through research core funding no. P6-0411, project J6-2581 (CANDAS - Computer-assisted multilingual news discourse analysis with contextual embeddings), and the young researcher grant.

References

- [Alnafesah et al.2020] Ghadi Alnafesah, Harish Tayyar Madabushi, and Mark Lee. 2020. Augmenting neural metaphor detection with concreteness. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 204–210, July.
- [Antloga and Donaj2022] Špela Antloga and Gregor Donaj. 2022. Corpus of metaphorical expressions in spoken Slovene language G-KOMET 1.0. Slovenian language resource repository CLARIN.SI.
- [Antloga2020] Špela Antloga. 2020. Korpus metafor KOMET 1.0. In *Proceedings of the Conference on Language Technologies and Digital Humanities*, pages 167–170.
- [Badryzlova et al.2022] Yulia Badryzlova, Olga Lyashevskaya, and Anastasia Nikiforova. 2022. Automated metaphor identification in Russian and its implications for metaphor studies. In *Distributed Computing and Artificial Intelligence, Volume 2: Special Sessions 18th International Conference*, pages 86–96.

- [Beigman Klebanov et al.2014] Beata Beigman Klebanov, Ben Leong, Michael Heilman, and Michael Flor. 2014. Different texts, same metaphors: Unigrams and beyond. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 11–17, June.
- [BNC Consortium2007] BNC Consortium. 2007. British national corpus, XML edition. Oxford Text Archive.
- [Choi et al.2021] Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MeIBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773.
- [Conneau et al.2020] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.
- [Cox1958] D. R. Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242.
- [Devlin et al.2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- [Do Dinh and Gurevych2016] Erik-Lân Do Dinh and Iryna Gurevych. 2016. Token-level metaphor detection using neural networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 28–33, June.
- [Florou et al.2018] Eirini Florou, Konstantinos Perifanos, and Dionysis Goutsos. 2018. Neural embeddings for metaphor detection in a corpus of greek texts. In *2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–4.
- [Hochreiter and Schmidhuber1997] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780, nov.
- [Krennmayr and Steen2017] Tina Krennmayr and Gerard J. Steen, 2017. *VU Amsterdam Metaphor Corpus*, pages 1053–1071. Springer.
- [Kutuzov et al.2018] Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397.
- [Lafferty et al.2001] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- [LeCun et al.2010] Yann LeCun, Koray Kavukcuoglu, and Clement F. Farabet. 2010. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 253–256.
- [Leong et al.2018] Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, June.
- [Leong et al.2020] Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29.
- [Lu and Wang2017] Xiaofei Lu and Ben Pin-Yun Wang. 2017. Towards a metaphor-annotated corpus of Mandarin Chinese. *Language Resources and Evaluation*, 51(3):663–694.
- [Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26.

- [Nguyen et al.2021] Minh Van Nguyen, Viet Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.
- [Prabhakaran et al.2021] Vinodkumar Prabhakaran, Marek Rei, and Ekaterina Shutova. 2021. How metaphors impact political discourse: A large-scale topic-agnostic study using neural metaphor detection. *Proceedings of the International AAI Conference on Web and Social Media*, 15(1):503–512.
- [Pramanick et al.2018] Malay Pramanick, Ashim Gupta, and Pabitra Mitra. 2018. An LSTM-CRF based approach to token-level metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 67–75, June.
- [Qimeng et al.2021] Yang Qimeng, Yu Long, Tian Shengwei, and Song Jinmiao. 2021. Uyghur metaphor detection via considering emotional consistency. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 895–905.
- [Saakyan et al.2022] Arkadiy Saakyan, Tuhin Chakrabarty, Debanjan Ghosh, and Smaranda Muresan. 2022. A report on the FigLang 2022 shared task on understanding figurative language. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 178–183, December.
- [Sanchez-Bayona and Agerri2022] Elisa Sanchez-Bayona and Rodrigo Agerri. 2022. Leveraging a new Spanish corpus for multilingual and cross-lingual metaphor detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 228–240, December.
- [Song et al.2021] Wei Song, Jingjin Guo, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021. A knowledge graph embedding approach for metaphor processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:406–420.
- [Steen et al.2010] Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins.
- [Stowe et al.2021] Kevin Stowe, Nils Beck, and Iryna Gurevych. 2021. Exploring metaphoric paraphrase generation. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 323–336, November.
- [Strapparava2018] Carlo Strapparava. 2018. Metaphor: A Computational Perspective. *Computational Linguistics*, 44(1):191–192, 03.
- [Tsvetkov et al.2013] Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51, June.
- [Turney et al.2011] Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690.
- [Ulčar and Robnik-Šikonja2021] Matej Ulčar and Marko Robnik-Šikonja. 2021. Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0. Slovenian language resource repository CLARIN.SI.
- [Ulčar and Robnik-Šikonja2020] Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT: Less is more in multilingual models. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020*, page 104–111.
- [Vaswani et al.2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.
- [Verdonik et al.2013] Darinka Verdonik, Iztok Kosem, Ana Zwitter Vitez, Simon Krek, and Marko Stabej. 2013. Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. *Language Resources and Evaluation*, 47(4):1031–1048.
- [Verdonik et al.2020] Darinka Verdonik, Sandi Majninger, Kaja Dobrovoljc, Špela Antloga, Aleksandra Zögling Markuš, Ines Voršič, Melita Zemljak Jontes, Melita Koletnik, Alenka Valh Lopert, Polonca Šek, Iztok Kosem, Majhenič Simona, and Ferme Marko. 2020. Korpus mladinske književnosti MAKŠ.

- [Wilcoxon et al.1970] Frank Wilcoxon, SK Katti, and Roberta A Wilcox. 1970. Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test. *Selected tables in mathematical statistics*, 1:171–259.
- [Zwitter Vitez et al.2022] Ana Zwitter Vitez, Mojca Brglez, Marko Robnik Šikonja, Tadej Škvorc, Andreja Vezovnik, and Senja Pollak. 2022. Extracting and analysing metaphors in migration media discourse: towards a metaphor annotation scheme. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2430–2439.

Evaluation of the Archivio Vi.Vo Architecture: A Case Study on the Reuse of Legacy Data for Linguistic Purposes

Roberta Bianca Luzietti

University of Pisa, Italy

roberta.luzietti@phd.unipi.it; robertabianca.luzietti@ilc.cnr.it

Abstract

The object of this paper is the evaluation of the Archivio Vi.Vo. architecture, developed within the CLARIN-IT consortium for the preservation and accessible consultation of historical oral archives. Following the first case study employing the Caterina Bueno archive, the goal is to now show how this innovative architecture is also suitable for conducting research investigations on the archival data and hosting a different type of archive. The real use case study presented in this contribution, aims at employing the Angela Spinelli archive for conducting a sociophonetic investigation on Tuscan vernacular.

1 Introduction

Thanks to infrastructures such as CLARIN (Common Language Resource and Technology Infrastructure) ERIC, different kinds of data and metadata can be safely stored in federated access repositories to become searchable, accessible in a digital format, standard compliant, interoperable with different tools and software and, most importantly, reusable (Krauwert and Hinrichs, 2014). To reach these goals the CLARIN consortia are committed to providing and making linguistic resources and tools available for social science research. In this regard, the Italian node CLARIN-IT recently developed the innovative Archivio Vi.Vo. architecture (Calamai et al., 2022) for the preservation, access, dissemination and reuse of historical audio archives. The architecture was conceived as a model for safeguarding the digitized content of analog carriers (preservation copies) and to make use of their recorded material (archival units). The internal structure of Archivio Vi.Vo. was specifically designed to make archival data interoperable, compliant with the CLARIN-IT infrastructure and safely deposited in the CLARIN repository. Along with reuse and dissemination, another reason for developing Archivio Vi.Vo. was the necessity to ensure the accountability of previous projects that have been dealing with the digitization and restoration of historical archives, such as Gra.fo (Calamai et al., 2013) (Calamai and Biliotti, 2017), and which risk becoming inaccessible once web portals remain unmaintained. Gra.fo was a two year project jointly conducted by the Scuola Normale di Pisa and the University of Siena (Regione Toscana PAR FAS 2007–2013) with the purpose to discover, digitize, catalogue and partially transcribe oral documents (e.g., oral biographies, ethno-texts, linguistic questionnaires, oral literature, etc.) collected within the Tuscan territory. The aim was to provide first-hand documentation of Tuscan speech varieties and Tuscan oral documents from the 1960s to the present. In the end, the project digitized nearly 3,000 hours of speech recordings stemming from around 30 oral archives collected by scholars and amateurs in the Tuscan territory.

The first Archivio Vi.Vo. case study was conducted on the Caterina Bueno archive characterized by a complex archival history and containing highly heterogeneous audio material such as ethnomusicological data (Calamai et al., 2022). The idea is now to validate the platform for hosting different types of archives, such as oral history archives, and see whether an archive collected within another field of study can still fit inside Archivio Vi.Vo. This paper presents a new case study involving the reuse of the Angela Spinelli archive by first including the data in Archivio Vi.Vo and then optimizing the exploration of

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

its content for conducting sociophonetic residual investigation on Tuscan vernacular, directly from the platform interface. The structure of the paper is as follows. Section 2 describes the Angela Spinelli archive. Section 3 illustrates the steps for including a new archive within the Archivio Vi.Vo architecture. Section 4 presents the benefits that come from using the platform for research purposes. Finally, section 5 presents the author's conclusions and future perspectives.

2 The Angela Spinelli Archive

The archive was collected by the historian Angela Spinelli at the beginning of the 1980s. The researcher conducted a series of face-to-face interviews with the inhabitants of small towns around the rural area of Prato along the upper Bisenzio river valley (e.g., Cantagallo, Vaiano, Montemurlo)¹ to ensure the preservation of memories about the historical events happening during the post Second World War period (Spinelli, 1981). The archive is made of 59 audio cassettes (corresponding to more than 120 hours of recording) and a large set of accompanying handwritten material consisting of important research protocol information and annotations made by the researcher during her work (Figure 1). The archive was legally acquired from Angela Spinelli and digitized in 2011 within the Gra.fo project (Calamai and Biliotti, 2017). Since the archive is unfortunately currently unavailable for online consultation from the project website, the Spinelli archive was recently stored into the Archivio Vi.Vo. architecture.

SEN	N. CATALOGO GENERALE	ENTE PROMOTORE	REGIONE	N.	DATA	RILEVAMENTO	RILEVATORE E SUA QUALIFICA	LOCALITA' RILEV.
10101	A)	PROVINCIA, COMUNE, FRAZIONE	TOSCANA	7	7-7-82		ANGELA SPINELLI - IES Scuola Sup. FI	MIGLIANA (CANTAGALLO)
	B)	OGGETTO proverbio x favola canzone modo di dire indovinello filastrocca						
	C)	CIRCOSTANZA(E) IN CUI E' PRODOTTO luogo(i) domestica - località Migliana - emittente genitori - I vecchi - persone presenti - I figli - gli amici - destinatario(i) - I figli - gli amici dato(i) temporale(±) ventennio tra le due guerre						
	D)	MODALITA' DI APPRENDIMENTO località e luogo(i) Migliana - ambiente domestico persona(e) emittente(i) genitori - i vecchi - età di apprendimento infanzia -						
	E)	VARIAZIONI DELL'OGGETTO località luogo persona(e) presente(i) destinatario(i) emittente(i) dato(i) temporale(i)						
	F)	DENOMINAZIONE DELL'OGGETTO NARRATO/DETTO/CANTATO 1) Chi taglia taglia, chi cuce ragguaglia - 2) L'uomo salvatico si rallegrava al cattivo - 3) Il maligno davanti ti liscia, di dietro ti graffia - 4) Che 'pittima' (pettegola): ha una lingua che taglia e cuce -						
	G)	TIPOLOGIA DELLA DENOMINAZIONE antropologica x sociale vita/cultura materiale x etico-morale religiosa economica						
	H)	SENSO TRASLATO O METAFORICO: TIPOLOGIA sociale economica etico morale x vita cultura materiale religioso						
	I)	LIVELLO(I) dell'OGGETTO attori principali attori secondari dato(i) socio-antropologico(i), culturale(i) degli attori principali secondari lo spazio il tempo la favola l'intreccio il ritmo la gestualità altre note						

Figure 1: Accompanying material from the Angela Spinelli Archive.

Angela Spinelli conducted her research under the supervision of Professor Roger Absalom (Absalom and Spinelli, 1998). Her method of investigation belongs to the field of oral history and consisted of eliciting unedited life histories of rural people born between 1880 and 1930 and belonging to different economic groups (e.g., small owners, sharecroppers, tenant farmers, charcoal burners, coal merchants, shepherds). Her research interest was the memory of everything pertaining to the aspects of the rural/material and associative life between the two world wars. Furthermore, she wanted to verify the essence of the long-term peasant mentality², at that time, challenged by the development model of small and medium-sized industries. To do so, she analyzed the periods of crisis (war) and transition (post-war)

¹Note that at the time of the interviews the area was still pertaining to the province of Florence, whereas from 1992 it became under the Prato jurisdiction.

²By long-term mentality she meant the type of peasant mentality for which changes are much slower than the artisanal and worker mentality which have more immediate transitions.

in which the old (rural) mentality was being tested against the new (industrial) one characterized by better job opportunities (in terms of both salary and physical efforts)(Spinelli, 1988). Another interesting aspect of the work of Angela Spinelli, is the use of keywords when interacting with the witnesses, to elicit their memory about the British Second World War refugees and other historical events.

The speakers that participated to the investigation were selected according to their age, origin and other speaker's recommendations. The interview protocol consisted in conducting three days of interviews and one day of pause in which Angela Spinelli would "close the ring": listen to the recordings, recapitulate her work and decide with whom and where to move forward with the interviews (Andreini and Clemente, 2007). The selected subjects could either be interviewed individually or in pairs (e.g., mother and son, husband and wife). The interviews started with the collection of personal details and information on the family and social network of the speakers as a way to put them at ease. From that point, the research would invite the speakers to discuss topics, such as: literacy, their relationship with British/American/South African allies, religion (ceremonies and pilgrimages), fascism, the female condition (working as servants, nurses, nuns), farming (the work cycles of coal, wheat, chestnuts, grape harvesting, work tools), diseases, popular culture (proverbial system, legends, superstitions), dowry and female hereditary succession, emigration, the economy, the family structure, life experiences during the first and second world war, marriage systems, land ownership, the resistance period after the Second World War, and "inurbation" phenomenon characterised by people abandoning the rural life for a better job and lifestyle in textile factories and railway construction sites in the nearby city of Prato. As shown in this section, the Spinelli archive is remarkably rich in both the number of hours of recording and the quality of oral history interviews, and it is, therefore, indeed well suited for research reuse purposes.

3 From Preservation Copy to Archival Unit in Archivio Vi.Vo.

The preservation process of a historical archive in the Archivio Vi.Vo. architecture begins with the upload of the digitized archival data in the storage server, followed by the creation of preservation copies within the platform (Calamai et al., 2022). Although the procedure for including archives in Archivio Vi.Vo. is still under development, it is possible to pre-announce that a single-sign-on (SSO) security access is currently being installed to provide different types of access to the archives inside the platform, according to the user and license.

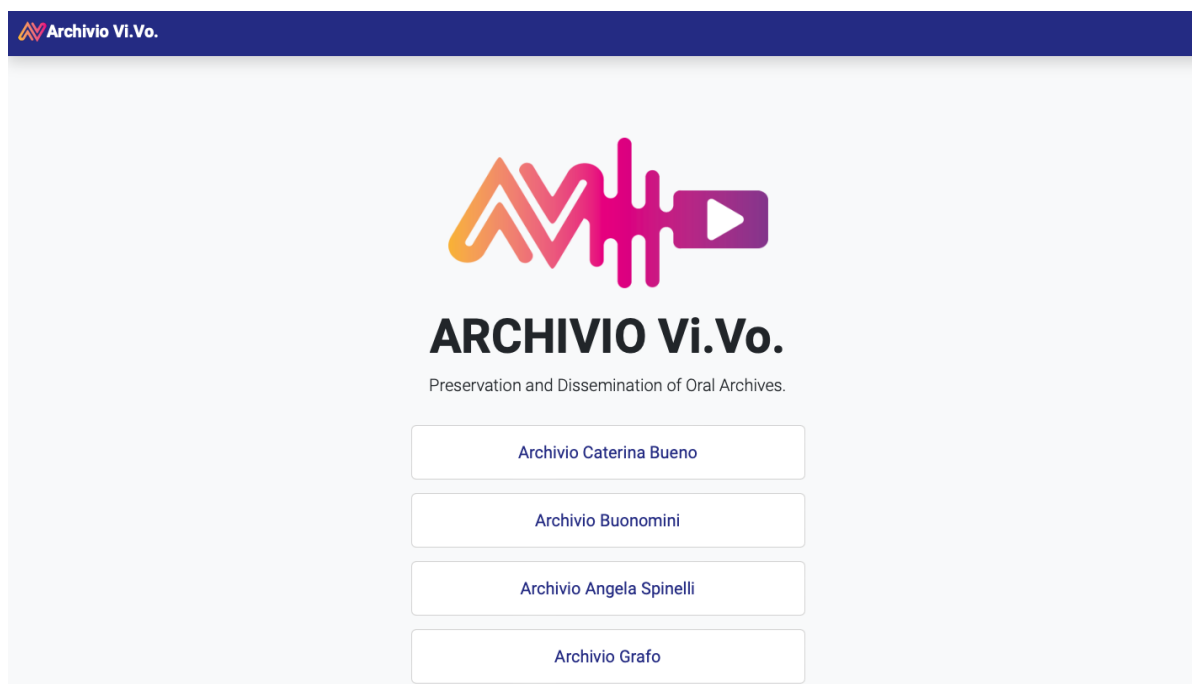


Figure 2: Archivio Vi.Vo. homepage.

The creation of the preservation copies consists in adding the metadata³ about the digitized copy of the carriers in Archivio Vi.Vo. The mandatory information is: the *denomination of the preservation copy* and the *digitization date*. The *creation date* is automatically detected by the system, whereas additional information, in this section, can include the *names of the people responsible for digitization* and of the *preservation copy supervisor(s)*. The required technical information (regarding the original carrier) is: the *signature(s) in the original archive*, the *owner*, the *container brand and model*, the *flange brand and model*, the *tape brand and model*, the *tape width*, the *carrier type*, the *carrier conditions before digitization*, the *restoration operations conducted before digitization*, and additional miscellaneous information (Figure 3). At the end of this process, it is also possible to add photographic material about the carrier (Figure 4).

Figure 3: Preservation copy metadata.

The status of completion of the preservation copies is indicated by a yellow (in progress) or green (completed) dot next to each file. The process of creation can be paused and resumed until the final confirmation after which it is no longer be editable. Most importantly, the deletion of a preservation copy created in Archivio Vi.Vo. does not entail the deletion of the file from the storage server. Between the preservation copies creation and the audio files conversion into the high quality compressed FLAC format, the architecture will soon provide access to the restoration interface (Calamai et al., 2022).


³In Archivio Vi.Vo., a customised set of metadata has been defined for the description of the preservation copy, inspired by other international standards for audio material description (IASA Technical Committee, 2009). The project adopted ISAD(G) and ISAAR standards for the archival units which have been interoperable with the CLARIN VLO infrastructure component which is part of CLARIN's Component Metadata Infrastructure (CMDI) (Calamai et al., 2020).

Archivio Vi.Vo. Archivio Angela Spinelli

Preservation Copies / LZRRN55a

LZRRN55a

Preservation Copy archived by n.d. on 06/02/2023 Delete



Digitization

Supervisor: n.d.
 Date: 18/07/2012

Carrier Description

Signature in the original archive: Angela Spinelli
 Other signature: n.d.
 Owner: n.d.
 Carrier Type: audiocassette
 Container Brand or Model: AmpeX ELN-90
 Flange Brand or Model: n.d.
 Tape Brand or Model: Magnetic
 Tape Width: 0.15 inch (3.81 mm)
 Carrier conditions before digitization: n.d.
 Restoration operations before digitization: n.d.
 Transcription of the carrier information: n.d.
 Notes: n.d.

Other Files

LZRRN55_01.jpg
 OriginalFiles/LZRRN55_01.jpg
 Notes: n.d.

Audio Description

Level of description	Action	Status
GROUP LZRRN55	Continue 	● In-Progress

Figure 4: Access interface for content exploration.

The Archivio Vi.Vo. section that follows is called description and is dedicated for accessing the content of the preservation copies and creating the archival units. This step is necessary to make the contents of the archive searchable and reusable. As explained in (Calamai et al., 2022) selecting the *clips* from audio file is needed because multiple communication events (such as conferences, interviews, concerts etc.) might have been recorded on a single carrier. Conversely, a single communicative event might have been recorded on more than one carrier. The communicative events consist in the various possible contents that can be found inside the recordings. In the description interface the audio can be played following the order in which it was recorded. The selection of the *clips* is done selecting the boundaries of beginning and end of a communicative event either using the slider or manually entering the time length. Once the *clips* are correctly segmented, descriptive annotations can be added to create the *regesto*⁴ (archival record) in alignment with the audio. For example, they can correspond to the various topics covered during an interview (Figure 5).

⁴Written summarization of an archival unit ((Calamai et al., 2022).

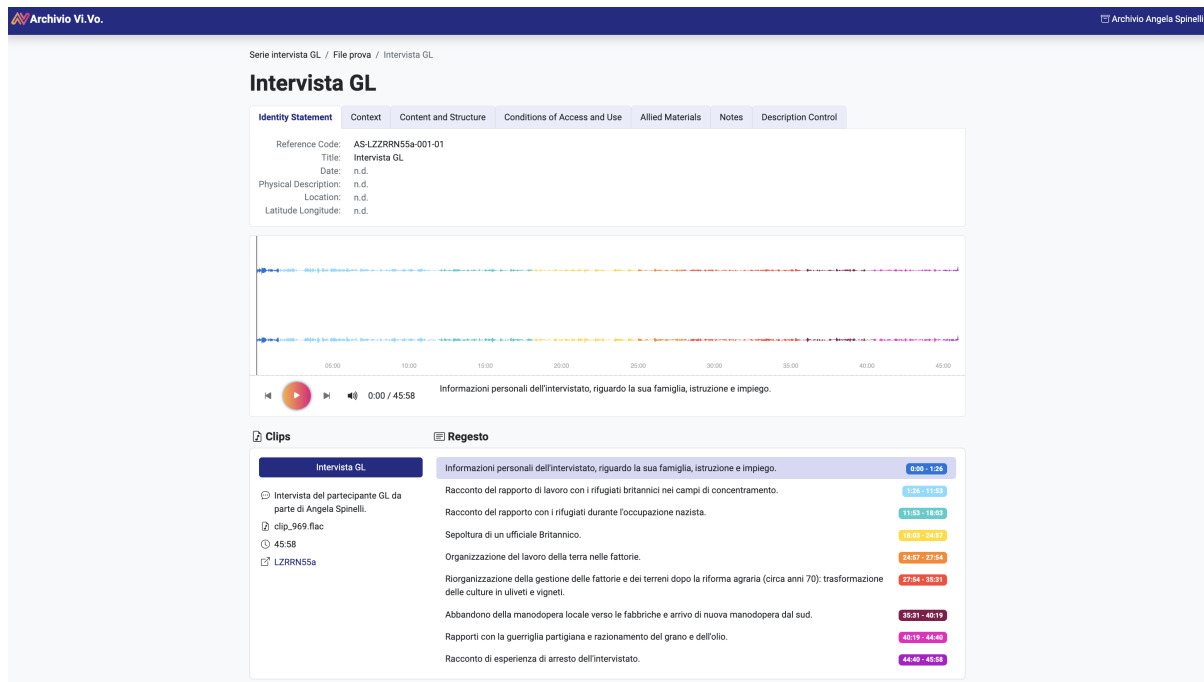


Figure 5: Access interface for content exploration.

The final result consists in a data structure in which the various contents of each preservation copy are neatly segmented and described, also known archival units, and can finally be safely employed for different purposes.

4 The Case Study

There are many benefits to reusing data, such as replicating studies, validate research claims, and achieving new discoveries, even across different disciplines. In the case of legacy data⁵, many scholars in the social sciences domain have shown how reusing (oral) historical archives can help the validation of research claims and the investigation of past, rare or disappeared phenomena that are no longer elicitable. Some examples of reuse of historical archives collected for different research purposes in linguistic investigations are: (De Fina, 2000), (Bornat, 2003), (Van De Mierop, 2009), (Schiffrin, 2009), (Roller, 2015), (Braber and Davies, 2016), (Renwick and Olsen, 2017) and (Nodari and Calamai, 2021).

As mentioned in section 1, employing the Archivio Vi.Vo. architecture can bring benefit to research projects on archival data. In this case the aim was to explore the content of the Angela Spinelli archive to investigate the presence of a residual phonetic phenomenon in Tuscan vernacular speech and, similarly to (Nodari and Calamai, 2021), to find correlations with social variables characterizing the speakers. The phenomenon under consideration is the phonetic reduction of the double (geminate) consonant /r:/ (rr) into a single /r/ in intervocalic contexts. The analysis is carried out on a set of words containing /r:/ in intervocalic position that also happen (but are not limited) to coincide with the keywords and topics addressed by the researcher during the interviews (e.g., *terra* (land) > *tera*, *guerra* (war) > *guera*). In past research on the description of Italian dialects along the La Spezia-Rimini isogloss, some authors reported the presence of rhotic degemination phenomena in the nearby territories investigated by Angela Spinelli (Giannelli, 1976) (Rohlf, 1966). However, until recent times the phenomenon has been under-investigated, especially from a (socio)phonetic perspective, in the Italian and Tuscan scientific literature (Celata et al., 2019). To find evidence of rhotic degemination in a phonetic laboratory would now be impossible since this phenomenon embodies the residual vestige of a past tendency challenged by the diffusion of more standard-like pronunciations involving the maintenance of the singleton–geminate

⁵“Data stored in obsolete audio media by individual researchers outside of archival sites such as libraries or data centres.” (Galatà and Calamai, 2019)

contrast. Hence, the necessity to look for degemination phenomena within oral history archives. As in (Nodari and Calamai, 2021), the assumption is that the consultation of past oral data collected within the area, where participants narrate emotional events, such as war or life-threatening situations and during which their speech is potentially less controlled and more spontaneous, could favor the presence of past non-standard forms. This is because during such interview settings the speakers focus more on what is being said rather than how (Labov, 1963).

Once the data were secured, the recordings identified, and described according to the various topics, with the help of the accompanying material (Monachini et al., 2021), the consultation of the interviews through the *regesto* was rapid and accurate. For example, having to look for keywords such as *terra* (land) and *guerra* it was decided to trace back the parts of interviews where the speakers would discuss war events and land cultivation, where they would have had higher chances to appear. Then, to verify the presence of the keywords within the selected interview portions, the corresponding audio was listened to while taking notes on whether the keywords were or were not present. Subsequently, the parts of interviews containing the keywords required transcription and phonetic annotation, and were the only two operations that could not be carried out from the platform. For the purpose of the research, the transcription was carried out manually and only for the interested discourse segments, whereas the data imported in PRAAT (Boersma, 2001) was manually traced back and extracted from the original preservation copy files. Correlations between the phenomenon and social variables of the speakers were, once again, traced back by listening to the audio and consulting the *regesto*. In this case, the sections to look for were those where the speakers would, for example, present themselves and talk about their family. The main advantage Archivio Vi.Vo. brought to this case study investigation was the optimization of the research protocol by reducing the time required for consultation of the data. Without the support of the architecture the whole process, involving listening to the recordings, operating the transcription and only then working on the data, would have been much more time consuming and dispersive.

Regarding the results of this research, it is important to note that the study is still ongoing and comprehensive results are not yet available. However, it could be said that historical oral archives can provide insightful perspectives for sociophonetic research. In particular, a preliminary analysis on part of the occurrences found in the Spinelli archive seems to indicate the presence of degemination in the Prato area and also suggests correlations with social and demographic variables. These findings offer a promising indication of the potential significance of this research, and highlight the importance of further investigation to fully understand the dynamics of degemination in this context. Overall, the ongoing research shows great potential for contributing to the understanding of sociophonetic variation in this area, and has the potential to yield valuable insights into past language usage and its relation to social and historical factors.

5 Discussion and Conclusion

The object of this article was to validate the Archivio Vi.Vo. architecture suitability for preserving different types of archives as well as favoring their reuse for different purposes. After the first Archivio Vi.Vo. case study involving the Caterina Bueno archive, the architecture requires validation for hosting different types of archives and support research projects. This paper presented a second case study involving the reuse of the Angela Spinelli archive to include the data in Archivio Vi.Vo and then optimize the exploration of its content for conducting sociophonetic residual investigation on Tuscan vernacular, directly from the platform interface. In addition, the data is finally be available for consultation and allows for the replication of the analysis, which is exceedingly rare in phonetics research.

What this case study shows is that the platform is, indeed, suitable for hosting a different type of (oral) archive and related accompanying material as well as help the optimization research protocols for what concerns data consultation. The main feature that emerged in this work and makes the platform a valuable research tool is its ability to provide a single environment for all the relevant archive material. This reduces the risk of material dispersion and facilitates the search for correlations between oral data and accompanying material. In the case study presented here, both of these aspects allowed for a quick and accurate retracing of the various details of different oral narratives, such as for reconstruction of the so-

cial network of different speakers. The system does not (yet) provide the ability to extract statistical data information about the data from the platform, but will be taken into consideration for further developments. However, choosing what kind of statistical variables to collect from the platform for quantitative analysis strongly depends on the research goals. In this case, one of the variables of interest was the duration (in milliseconds) of the consonants and preceding vowels that could not be directly extracted and measured directly from the architecture and required the use of PRAAT. As a first step towards linking external tools to Archivio Vi.Vo. could be the integration of already existing resources within CLARIN, such as the Transcription Portal for Interview Data (Draxler et al., 2020)) Given the benefits obtained in terms of saving time and resources with Archivio Vi.Vo, it is possible to conclude that the platform is certainly suitable for supporting research investigations dealing with oral archives, without excluding the possibility of implementing new functions in the future. As more (and different types of) archives are uploaded to the platform, in addition to allowing for more validation examples, it is expected that it will potentially give rise to the implementation of additional features that will allow the use of Archivio Vi.Vo. as a research tool to be expanded even further.

Future perspectives regarding Archivio Vi.Vo. concern: i) incorporating the SSO federated access to the platform, ii) completing the restoration interface, iii) implementing the possibility to export audio data directly from the platform into software for data transcription and for phonetic analysis and annotation, and iv) promote the use of the platform within the CLARIN and CLARIN-IT consortia training activities.

References

- Roger Absalom and Angela Spinelli. 1998. Contadini e gli ex prigionieri anglo-sassoni a prato nella seconda guerra mondiale. in *Giacomo Becattini (a cura di), Prato Storia di una città. Il modello pratese (1943 - ad oggi) Firenze, Le Monnier IV*, pages 43–81.
- Alessandro Andreini and Pietro Clemente. 2007. I custodi delle voci. *Archivi orali in Toscana: primo censimento, Firenze, Tipografia Regionale*.
- Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glott International* 5:9/10, pages 341–345.
- Joanna Bornat. 2003. A second take: Revisiting interviews with a different purpose. *Oral History*, 31(1):47–53.
- Natalie Braber and Diane Davies. 2016. Using and creating oral history in dialect research. *Oral History*, pages 98–107.
- Silvia Calamai and Francesca Biliotti. 2017. The gra. fo project: from collection to dissemination. *Umanistica Digitale*.
- Silvia Calamai, Francesca Biliotti, Pier Marco Bertinetto, Chiara Bertini, Irene Ricci, and Gianfranco Scuotri. 2013. The gra. fo sound archive: Architecture, methods and purpose. In *Proceedings of the 2013 Digital Heritage International Congress (DigitalHeritage)*, page 439.
- Silvia Calamai, Niccolò Pretto, Monica Monachini, Maria Francesca Stamuli, Silvia Bianchi, Pierangelo Bonazoli, and Unione dei Comuni Montani. 2020. Building a home for italian audio archives. In *proceedings of CLARIN Annual Conference*.
- Silvia Calamai, Duccio Piccardi, Niccolò Pretto, Giovanni Candeo, Maria Francesca Stamuli, and Monica Monachini. 2022. Not just paper: Enhancement of archive cultural heritage. *CLARIN The Infrastructure for Language Resources, ISBN 9783110767377, Fišer D. and Witt A. (eds.), published by Walter De Gruyter and Co (Berlin, DEU), vol. 1*.
- Chiara Celata, Alessandro Vietti, and Lorenzo Spreafico. 2019. An articulatory account of rhotic variation in tuscan italian: Synchronized uti and epg data. *OXFORD LINGUISTICS*, pages 91–117.
- Anna De Fina. 2000. Orientation in immigrant narratives: The role of ethnicity in the identification of characters. *Discourse Studies*, 2(2):131–157.
- Christoph Draxler, Henk van den Heuvel, Arjan van Hessen, Silvia Calamai, and Louise Corti. 2020. A clarin transcription portal for interview data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3353–3359.

- Vincenzo Galatà and Silvia Calamai. 2019. Looking for hidden speech archives in Italian institutions. In *CLARIN Annual Conference 2018*, page 104.
- Luciano Giannelli. 1976. *Toscana*. Pisa, Pacini.
- Steven Krauwer and Erhard Hinrichs. 2014. The CLARIN research infrastructure: resources and tools for e-humanities scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1525–1531. European Language Resources Association (ELRA).
- William Labov. 1963. The social motivation of a sound change. *Word*, 19(3):273–309.
- Monica Monachini, Maria Francesca Stamuli, Silvia Calamai, Niccolò Pretto, and Silvia Bianchi. 2021. The grey-side of audio archives. *The GL-conference series. Conference proceedings, ISSN 1386-2316, published by TransAtlantic (Amsterdam, Paesi Bassi)*, vol. 22:34–37.
- Rosalba Nodari and Silvia Calamai. 2021. Degemination in marginal Tuscan speech: temporal analysis in legacy speech data. In *D. Recasens, F. Sánchez-Miret (curated by) Sound change in Romance: phonetic and phonological issues (pp. 67-85)*. München : Lincom.
- Margaret EL Renwick and Rachel M Olsen. 2017. Analyzing dialect variation in historical speech corpora. *The Journal of the Acoustical Society of America*, 142(1):406–421.
- Gerald Rohlfs. 1966. *Grammatica storica della lingua italiana e dei suoi dialetti (Fonetica)*. Torino: Einaudi.
- Katja Roller. 2015. Towards the 'oral' in oral history: using historical narratives in linguistics. *Oral History*, pages 73–84.
- Deborah Schiffrin. 2009. Crossing boundaries: The nexus of time, space, person, and place in narrative. *Language in Society*, 38(4):421–445.
- Angela Spinelli. 1981. Le comunità contadine del pratese nella lotta di liberazione e nell'assistenza ai prigionieri britannici evasi 1943-45. fonti orali e ricerca storica nell'indagine su una classe subalterna. *Argomenti storici*, 8, pages 1–27.
- Angela Spinelli. 1988. Archivio sonoro delle comunità contadine dell'alta val Bisenzio. *Rassegna degli Archivi di Stato*, 48, 1-2, pages 232–238.
- Dorien Van De Mierop. 2009. Exploring the influence of the interview as a research method on the construction of identities in narrative. In *International Symposium on Theoretical & Applied Linguistics, Date: 2009/04/06-2009/04/05, Location: Thessaloniki, Greece*.

It-Sr-NER: CLARIN Compatible NER and Geoparsing Web Services for Italian and Serbian Parallel Text

Olja Perišić

University of Turin, Italy
olja.perisic@unito.it

Ranka Stanković

University of Belgrade, Serbia
ranka.stankovic@rgf.bg.ac.rs

Milica Ikonić Nešić

University of Belgrade, Serbia
milica.ikonik.nesic@fil.bg.ac.rs

Mihailo Škorić

University of Belgrade, Serbia
mihailo.skoric@rgf.rs

Abstract

The paper will showcase the outcomes of the "It-Sr-NER: Web services for named entities recognition, linking and mapping" project for Serbian and Italian languages. The project was a collaboration between the University of Turin and the Society for Language Resources and Technologies JeRTeh, with the goal of creating the It-Sr-NER web service. This service is designed to annotate named entities such as people, places, organizations, ethnicities, events, and works of art in text, and display them on a map.

1 Introduction

The main motivation for starting "It-Sr-NER: Web services for named entities recognition, linking and mapping" project was the lack of tools and resources for annotating, researching, and analyzing bilingually aligned Italian-Serbian texts. At the same time there is a significant absence of corpus tools in the teaching of foreign languages in Serbia, as noted in recent research (Vitaz and Poletanović, 2020). However, on an individual level and through personal initiatives in the teaching of Serbian as a foreign language in Italy and the Italian language in Serbia, it has been shown that corpora can be highly beneficial in many ways and that students are eager to use them in collaborative work and independent research (Moderc, 2015b; Perišić, 2021). One of the challenges in teaching the Serbian language to foreign students is the rich and sophisticated morphology, which includes declensions of toponyms and other named entities that can be difficult for students to recognize and reduce to their basic form. This is due to factors such as similar endings for the masculine and neuter gender in most grammatical cases, the presence of certain toponyms only in the plural form, the so-called *pluralia tantum* (Berane, Udine, etc.), phonetic transcriptions of foreign names, and some orthographic inconsistencies (Vitas and Lažetić-Pavlović, 2008).

A team of experts from the University of Turin and the Society for Language Resources and Technologies JeRTeh have partnered as part of CLARIN's call "Bridging Gaps" to develop web services for annotating named entities in text. These services ensure the linking of named entities with Wikidata and provide geoparsing, which includes geolocating recognized locations and displaying them on a map. These services specifically target names of persons, places, organizations, ethnicities, events, and works of art as named entities.

The primary objective of the project was to create and publish web applications and services for monolingual and bilingual parallel texts within the CLARIN infrastructure as well as on the platform of the Society for Language Resources and Technologies JeRTeh. The project also aimed to create and publish an Italian-Serbian corpus of 10,000 segments of extracted and aligned sentences, selected from classics of Italian and Serbian literature. The outcomes of the project are not restricted to the Italian-Serbian language combination, the developed services can be used to process texts in twenty-four different languages.

The project was initiated and led by Olja Perišić, a professor at the University of Turin, where she teaches the Serbian and Croatian language. On behalf of JeRTeh, the development of the services was

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

led by Professor Ranka Stanković in collaboration with Professor Duško Vitas. Further information about the project can be found on the website of the Society for Language Resources and Technologies JeRTeh, which includes the It-Sr-NER CLARIN compatible named entity recognition (NER), named entity linking (NEL) and geoparsing web services for parallel texts.¹

2 Italian-Serbian Parallel Corpus

Parallel corpora have been found to be a crucial tool in foreign language teaching as they help, at the beginner level, to acquire morphosyntax and lexis. Observing two or more languages in parallel facilitates contrastive analysis, which allows for the examination of similarities and differences in language structures by providing a large number of sentence examples in context. As Sinclair pointed out at an early stage of development of corpus linguistics: "The language looks rather different when you look at a lot of it at once" (Sinclair, 1991). At the intermediate and advanced level, parallel corpora are an effective tool in teaching translation as they facilitate disambiguation of word senses and definition of polysemic vocabulary, which are often underrepresented in bilingual dictionaries especially for this language combination (Moderc, 2015a; Perišić Arsić, 2018). A translation is always linked to the context of the target language, to the individual style of each translator and his/her interpretation of the original text. The possibility to compare different translations of a single text can highlight any ambiguities or inconsistencies already present in the source text. These ambiguities, due to several linguistic reasons, concern the register and are attributable to various cultural factors, but they are hardly noticed in the monolingual analysis of a text (Perišić, 2023). At the same time, it has been noted that there is a lack of representative parallel corpora even for major world languages (Granger, 2018).

To overcome this problem, as a first step in the project, it was necessary to create an Italian-Serbian corpus of 10,000 aligned segments (sentences) taken from ten different novels. The novels by Italian writers represented in the corpus are: Umberto Eco's "The Name of the Rose", Carlo Collodi's "The Adventures of Pinocchio", Elena Ferrante's "Those Who Leave and Those Who Stay", and Luigi Pirandello's "One, None and a Hundred Thousand". The corpus also includes five novels by Serbian writers: Ivo Andrić's "Legends of Anika" and "The Bridge on the Drina", Borisav Stanković's "Impure Blood", Branislav Nušić's "Municipal child: the novel of an infant", Danilo Kiš's "Garden, Ashes". Additionally, the corpus also includes Italian and Serbian translations of Jules Verne's "Around the World in Eighty Days" in order to support the main task of the project which is annotating named entities.

The novels were aligned and converted to TMX (Translation Memory eXchange) format using the ACIDE program, which is designed for creating parallel corpora (Obradović et al., 2008; Krstev and Vitas, 2011). Figure 1 on the left presents the samples of translation units, which contain the translation equivalents in tag <tuv>. The segments in Italian and Serbian are paired and numbered, with each segment indicating the language through the attribute "xml:lang". The ACIDE program not only creates the TMX document, but also generates an HTML representation, as shown in Figure 1 on the right.

The It-Sr-NER corpus² is available on the ILC4CLARIN B Center and can be accessed through the VLO (Virtual Language Observatory)³. The corpus, in a compressed format, includes the aligned bilingual version, as well as individual monolingual versions, and named entities that have been automatically tagged (as detailed in Section 3). The corpus and additional information can also be found in the Github⁴.

The corpus, which includes the complete novels from which the published version of 10,000 segments were extracted, is not only downloadable but is searchable on the Bibliša⁵ digital library. The left side of the Figure 2 presents browsing of documents (novels) with additional possibilities for authorised users for editing of metadata and aligned sentences (on the right).

Parallel corpora are useful for translation research, and the use of concordances in contrastive linguistics can improve the study of cross-linguistic phenomena. The resources developed in this project can

¹<https://jerteh.rs/index.php/it-sr-ner-3/>

²<http://hdl.handle.net/20.500.11752/OPEN-980>

³<https://vlo.clarin.eu/>

⁴<https://github.com/jerteh/It-Sr-NER/tree/main/corpus>

⁵<http://biblisha.jerteh.rs>

Italian (it)	Serbian (sr)
<p>n2 Lasciai il tavolo frastornata, stentavo a prendere atto che Nino era davvero lì, a Milano, in quella saletta.</p> <p>n3 Eppure eccolo, già mi veniva incontro sorridendo ma con passo controllato, senza fretta.</p>	<p>n2 Ustadoh od stola pometena, bilo mi je teško da ubedim samu sebe da je Nino zaista tu, u Milanu, u toj maloj sali.</p> <p>n3 Pa ipak, eto ga kako mi ide u susret, sa osmehom ali odmerena koraka, bez žurbe.</p>


```

<tu>
  <prop type="Domain"/>
  <tuv xml:lang="it" creationid="n2" creationdate="20211014T224355Z">
    <seg>Lasciai il tavolo frastornata, stentavo a prendere atto che Nino era davvero
    li, a Milano, in quella saletta.</seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n2" creationdate="20211014T224355Z">
    <seg>Ustadoh od stola pometena, bilo mi je teško da ubedim samu sebe da je Nino
    zaista tu, u Milanu, u toj maloj sali.</seg>
  </tuv>
</tu>
<tu>
  <prop type="Domain"/>
  <tuv xml:lang="it" creationid="n3" creationdate="20211014T224355Z">
    <seg>Eppure eccolo, già mi veniva incontro sorridendo ma con passo controllato,
    senza fretta.</seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n3" creationdate="20211014T224355Z">
    <seg>Pa ipak, eto ga kako mi ide u susret, sa osmehom ali odmerena koraka, bez
    žurbe.</seg>
  </tuv>
</tu>

```

Figure 1: Two aligned segments of translation equivalents, Italian and Serbian (HTML left, TMX right).

The screenshot shows the 'BIBLIŠA: ALIGNED COLLECTION SEARCH TOOL' interface. It features a navigation menu with options like 'Home', 'Metadata browse', 'Metadata search', 'Mongo search', 'Manage data', 'Help', 'Tutorial', and 'About'. Below the menu is a table of search results for 'ITSrKOR'. The table has columns for 'ID', 'EN' (English/Italian), 'SR' (Serbian), and 'SR' (actions: Delete, Edit). A red arrow points to the entry for ID n4, which shows the alignment between the Italian text 'Il libro, corredato da indicazioni storiche invero assai povere, asseriva di riprodurre fedelmente un manoscritto del Quattordicesimo secolo, a sua volta trovato nel monastero di Melk dal grande erudito secentesco, a cui tanto si deve per la storia dell'ordine benedettino.' and the Serbian text 'Knjiga, opremljena uistinu oskudnim istorijskim podacima, tvrdila je da verno prenosi jedan rukopis iz XIV veka koji je, opet, pronašao u manastiru u Melku veliki erudita XVII stoleća, onaj kome toliko dugujemo povodom istorije benediktinskog reda.'

ID	EN	SR	SR
n1	Umberto Eco, Il nome della rosa	Umberto Eko, Ime ruže	Delete Edit
n2	Naturalmente, un manoscritto.	Naravno, rukopis	Delete Edit
n3	Il 16 agosto 1968 mi fu messo tra le mani un libro dovuto alla penna di tale abate Vallet, "Le manuscrypt de Dom Adson de Melk, traduit en français d'après l'édition de Dom J. Mabillon" (Aux Presses de l'Abbaye de la Source, Paris, 1842).	Šesnaestog avgusta 1968. dospela mi je do ruku knjiga iz pera izvesnog opata Valea, Le manuscrypt de Dom Adson de Melk, traduit en franCais d'aprEs l'Édition de Dom J. Mabillon (Aux Presses de l'Abbaye de la Source, Paris, 1842).	Delete Edit
n4	Il libro, corredato da indicazioni storiche invero assai povere, asseriva di riprodurre fedelmente un manoscritto del Quattordicesimo secolo, a sua volta trovato nel monastero di Melk dal grande erudito secentesco, a cui tanto si deve per la storia dell'ordine benedettino.	Knjiga, opremljena uistinu oskudnim istorijskim podacima, tvrdila je da verno prenosi jedan rukopis iz XIV veka koji je, opet, pronašao u manastiru u Melku veliki erudita XVII stoleća, onaj kome toliko dugujemo povodom istorije benediktinskog reda.	Delete Edit
n5	La dotta trouvaille (mia, terza dunque nel tempo) mi rallegrava mentre mi trovavo a Praga in attesa di una persona cara.	Učeno otkriće (moje, dakle treće u vremenskom sledu) radovalo me je dok sam boravio u Pragu, iščekujući jednu dragu osobu.	Delete Edit
n6	Sei giorni dopo le truppe sovietiche invadevano la sventurata città.	Šest dana kasnije sovjetske trupe zaposale su zlosrećni grad.	Delete Edit
n7	Riuscivo fortunosamente a raggiungere la frontiera austriaca a Linz, di lì mi portavo a Vienna dove mi ricongiungevo con la persona attesa, e insieme risalivamo il corso del Danubio.	Preturivši štošta preko glave, dokopah se austrijske granice kod Linca, odande stigoh do Beča, gde mi se pridruži iščekivana osoba, pa zajedno krenusmo uz Dunav.	Delete Edit
n8	In un clima mentale di grande eccitazione leggevo, affascinato, la terribile storia di Adso da Melk, e tanto me ne lasciai assorbire che quasi di getto ne stesi una traduzione, su alcuni grandi quaderni della Papeterie Joseph Gibert, su cui è tanto piacevole scrivere se la penna è morbida.	U atmosferi velikog duševnog uzbuđenja čitao sam, sav očaran, strašnu povest Adsa iz Melka, i ona me je ophvala toliko da sam bezmalo u jednom dahu sačinio prevod, u nekoliko velikih svezaka koje pravi Papeterie Joseph Gibert i u kojima je takvo uživanje pisati ako je olovka meka.	Delete Edit

Figure 2: Aligned text from ItSrKor in Bibliša digital library.

be utilized by students of Italian language in Serbia and Serbian language in Italy as they are open and accessible to other students and researchers in tertiary and pre-tertiary education.

3 Web Services for Named Entity Recognition and Linking

The It-Sr-NER services ⁶, which are stored in the CLARIN repository, can process not only monolingual texts in 24 languages, but also bilingual texts (represented in the TMX format), and successfully annotate them.

⁶<http://hdl.handle.net/20.500.11752/OPEN-981>

The ultimate goal of the project was to integrate the developed web services into the European infrastructure for language resources and technologies, specifically the Language Resource Switchboard platform⁷. The initial aim to annotate named entities for Italian and Serbian was later extended to other languages for which models were available. These models, which were trained using the spaCy⁸ library, were downloaded for each language from the corresponding repository⁹ in order to incorporate them.

For Italian, the *it_core_news_sm3.4.0* model, which was trained on the automatically created corpus, *WikiNER*¹⁰, based on Wikipedia text and structure (Nothman et al., 2013), was used. For Serbian, a model trained on the *SrpCNNER* corpus of old Serbian novels (Šandrih Todorović et al., 2021) available on the European Language Grid (ELG) platform was used¹¹.

The University Library of Mannheim developed an open source system OpenTapioca¹², which links named entities to concepts in Wikidata (Delpeuch, 2019). By using the spaCy wrapper spaCyOpenTapioca¹³, the application can not only recognize and annotate named entities, but also link them with items in Wikidata. The final outcome is a web service that can display recognized named entities on a map.

Four types of web services have been developed: NER, NER+NEL, NEL and geoparsing. Further on, for each service type two services were developed: one for monolingual and one for bilingual resources.

- The NER (Named Entity Recognition) process uses trained language models from the spaCy library to recognize named entities based on the classes listed in Table 1.
- NER+NEL is an extension of the NER process. In addition to recognizing named entities, it also links the annotated entities with Wikidata when possible. This is achieved by using the functions of the spaCyOpenTapioca service, and is applied only to the recognized named entities, which are the text inside the XML tag.
- NEL (Named Entity Linking) is the process of recognizing and linking named entities with Wikidata, using the recognition capabilities of the spaCyOpenTapioca system. The recognized named entities are annotated with the tag `<WDT>` and the class of the named entity is identified using the label attribute.
- Geoparsing - using the *geopy* library¹⁴ for geolocating named entities of the *LOC* class that are present in wikidata, and then displaying them on a map using the *folium*¹⁵ library.

To standardize the labels used for different classes of named entities across language-specific models, unification of tagset is prepared as presented in Table 1. The *PERS* class label, which marks persons, was set as the default label to which corresponding labels from other models were mapped. Furthermore, the labels denoting locations and geopolitical entities have been unified to the *LOC* label, regardless of any other labels that may have been used (such as *GPE*, *LC*, *placeName*, or *geogName*).

The label *NORP* (nationalities or religious or political groups) for nationalities, political and religious groups from Japanese and Finnish models and *NAT_REL_POL* from Romanian model were mapped to the *DEMO* label, which denotes demonyms and ethnic relations (Stanković et al., 2021). This mapping is done consistently for all classes and can be found in the configuration file¹⁶. Since some language models have a more extensive set of named entity classes, for example, English has 18 classes and Romanian has 16, a column for ignored labels is defined in the configuration file.

⁷<https://switchboard.clarin.eu/tools>

⁸<https://spacy.io/>

⁹<https://spacy.io/models>

¹⁰https://figshare.com/articles/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500

¹¹<https://live.european-language-grid.eu/catalogue/id/9484>

¹²<https://opentapioca.org>

¹³<https://pypi.org/project/spacyopentapioca/>

¹⁴<https://pypi.org/project/geopy/>

¹⁵<https://python-visualization.github.io/folium/>

¹⁶https://github.com/jerteh/It-Sr-NER/blob/main/config/lng_config.csv

NER class tag	Description of the entity class	Mapped also
PERS	Names, surnames, nicknames and their combinations (of real people and fictional characters, including gods and saints).	PER (French, German, Italian, Portuguese, Spanish...), PRS (Swedish), PERSON (English, Finnish, Greek, Macedonian...), persNAME (Polish), PS (Korean)
LOC	Continents, countries, regions, settlements, oronyms, bodies of water, names of celestial bodies, city locations.	LOC+GPE (Chinese, English, Finnish, Greek, Macedonian, Romanian...), LC (Korean), placeName and geogName (Polish)
ORG	Names of companies, political parties, educational institutions, sports teams, hospitals, museums, libraries, hotels, cafes, churches and shrines.	ORGANIZATION (Romanian), orgName (Polish), ORG+GPE_ORG (Norwegian Bokmål), OG (Korean)
DEMO	Residents of countries, cities, regions or ethnic groups; derived adjectives from the name of the location.	NORP (Chinese Dutch, English, Finnish, Japanese, Macedonian), NAT_REL_POL (Romanian)
EVENT	Names of events that recur regularly or happened once but they have their own name: natural disasters, revolutions, battles, wars.	EVN (Swedish), EVT (Norwegian Bokmål)
WORK	Titles of books, plays, poems, paintings, sculptures, newspapers.	WORK_OF_ART (Romanian, Dutch, English, Japanese, Macedonian, Finnish), WRK (Swedish)

Table 1: Named entity classes.

Web services that use named entity linking (NER+NEL and NEL) provide additional information about named entities as *xml* attributes: the entity type (*label*), description (*desc*), and a link to the Wikidata knowledge base (*ref*), in addition to the classes already associated to the entities.

It was previously stated that the input can be either monolingual or bilingual text. For bilingual resources, the input must be in the form of a valid *TMX* document. The output of three services for bilingual resources is shown in Figure 3, where the first possibility is NER, the second is NER+NEL, and the third is NEL service, with *spacyOpenTapioca*-based services linking recognized named entities to items in Wikidata for both languages.

The program code, web services, web application, and parallel corpora from the project have all been released¹⁷ under open licenses, allowing for free use in research and commercial activities.

The primary development of the project took place from June to September 2022, with further adjustments made during the subsequent fine-tuning phase. In order to achieve all the results, the core team of four researchers received support from an additional three researchers. Adapting the web service to function with the CLARIN infrastructure presented a challenge, but thanks to the assistance of the CLARIN team, the verification and publication of the service were successful. The evaluation was conducted on a limited dataset for Serbian and Italian. Results revealed that PERS and ORG were better identified in Italian than in Serbian, while Serbian LOC performed better. The Italian model did not include DEMO, WORK, and EVENT. Overall, the evaluation indicates the need for further model improvement.

¹⁷<https://github.com/jerteh/It-Sr-NER>

```

<tu>
  <prop type="Domain"/>
  <tuv xml:lang="it" creationid="n45" creationdate="20220825T212333Z">
    <seg><PERS>Nino</PERS> spiegò che viveva a <LOC>Milano</LOC> da anni, si occupava di
    geografia economica, apparteneva - e sorrise - alla categoria più miserabile
    della piramide accademica, vale a dire gli assistenti.</seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n45" creationdate="20220825T212333Z">
    <seg><PERS>Nino</PERS> je objasnio da već godinama živi u <PERS>Milanu</PERS>, da se
    bavi ekonomskom geografijom, da pripada - i tu se osmehnu - najnižem staležu
    akademske piramide, takoreći asistentima.</seg>
  </tuv>
</tu>
<tu>
  <prop type="Domain"/>
  <tuv xml:lang="it" creationid="n45" creationdate="20220825T212333Z">
    <seg><PERS>Nino</PERS> spiegò che viveva a <LOC
      ref="https://www.wikidata.org/wiki/Q490" desc="major city in Italy"
    >Milano</LOC> da anni, si occupava di geografia economica, apparteneva - e
    sorrise - alla categoria più miserabile della piramide accademica, vale a dire
    gli assistenti.</seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n45" creationdate="20220825T212333Z">
    <seg><PERS>Nino</PERS> je objasnio da već godinama živi u <PERS
      ref="https://www.wikidata.org/wiki/Q490" desc="major city in Italy"
    >Milanu</PERS>, da se bavi ekonomskom geografijom, da pripada - i tu se
    osmehnu - najnižem staležu akademske piramide, takoreći asistentima.</seg>
  </tuv>
</tu>
<tu>
  <prop type="Domain"/>
  <tuv xml:lang="it" creationid="n45" creationdate="20220825T212333Z">
    <seg>Nino spiegò che viveva a <WDT ref="https://www.wikidata.org/wiki/Q490"
      label="LOC" desc="major city in Italy">Milano</WDT> da anni, si occupava di
    geografia economica, apparteneva - e sorrise - alla categoria più miserabile
    della piramide accademica, vale a dire gli assistenti.</seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n45" creationdate="20220825T212333Z">
    <seg>Nino je objasnio da već godinama živi u <WDT
      ref="https://www.wikidata.org/wiki/Q490" label="LOC"
      desc="major city in Italy">Milanu</WDT>, da se bavi ekonomskom geografijom,
    da pripada - i tu se osmehnu - najnižem staležu akademske piramide, takoreći
    asistentima.</seg>
  </tuv>
</tu>

```

Figure 3: NER, NER+NEL and NEL output for bilingual resources in TMX format.

4 Use Cases

The web services discussed in Section 3 can be accessed and utilised in a variety of ways. Figure 4 illustrates the web services integrated on the Language Resource Switchboard platform. Bilingual resources must be inputted as an XML file (in TMX format), while monolingual resources can be submitted as a text file or entered directly into the provided field on the web application form. The integration of the web service in the CLARIN infrastructure allows for greater visibility and accessibility for researchers and educators, and the ability to easily share resources and collaborate on projects. (de Jong et al., 2022; Draxler et al., 2022)

Figure 5 illustrates an example of the results of processing a bilingual text (submitted as a TMX document) on the CLARIN platform Language Resource Switchboard using the NER+NEL service. The output shows the processing results for both languages presented simultaneously, displaying the named entities recognized and linking them to knowledge base. Each named entity category is color-coded, in order to better visualize the results to the end user.

The figure also illustrates the capability of displaying the link to wikidata and description of an item (determiner), in this case, Florence (Q2044). It is shown that the recognized named entity Florence (*Firenze* in Italian) is associated with an underlined style, which is a feature of the web services. Users can hover over the underlined text to see the description of the item, this is an additional feature provided to help users understand the context and meaning of the named entities.

The web services described are also accessible via the web application at <https://ners.jerteh.rs>. This is an implementation of the previously mentioned web application (developed for the It-Sr-Ner project), with embedded API endpoints, that are targeted by the app instance and the Switchboard alike. Since

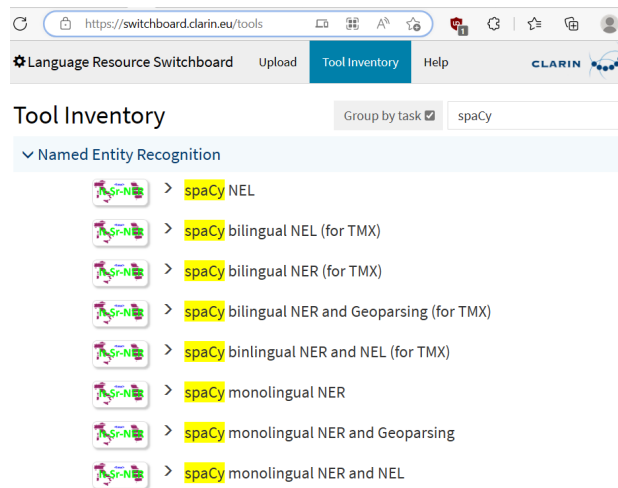


Figure 4: Presentation of the integrated web service on the CLARIN infrastructure.

PERS	LOC	ORG	DEMO	WORK	EVENT	Download XML
n75	Avrei trovato il modo di tirarmi addosso Nino con tutti gli anni che erano passati, dalle elementari al liceo, fino al tempo di Ischia e di piazza dei Martiri .					
n75	Našla bih način da privučem Nina k sebi nakon svih onih godina, od osnovne škole do gimnazije, pa sve do vremena provedenog na Iskiji i u radnji na Trgu mučenika.					
n116	Poi mi disse che aveva letto l'articolo di Sarratore , ma solo perché un fornitore si era dimenticato il Roma nel negozio.					
n116	Zatim mi reče da je pročitala Saratoreov članak, ali samo zato što neki od dobavljača beše zaboravio Romu u radnji.					
n117	Aveva saputo da mia madre che mi sarei sposata presto con un professore dell'università e che sarei andata a vivere a Firenze in Italian city, located in Tuscany .					
n117	Od moje majke beše saznala da ću se uskoro udati za jednog univerzitetskog profesora i da ću otići da živim u Firenci kao prava					
n256	E senti che non c'era continuità tra i tempi di Ischia e la fabbrica di salumi: in mezzo si stendeva il vuoto, e nel salto da uno spazio all'altro Bruno – forse perché il padre di recente era stato male e il peso dell'azienda (i debiti, qualcuno diceva) gli era caduto all'improvviso sulle spalle – si era guastato.					
n256	I oseti kako postoji prekid u vremenu između onog perioda na Iskiji i ovoga sada u fabrici salama, da se njim proteže bezdan, i da se Bruno iznenada – možda zato što mu je otac već neko vreme bio bolestan pa je čitav teret fabrike (beše načula nešto o dugovima) iz vedra neba pao na njegova pleća – nekako iskvario.					
n275	Pasquale, appena accennava alla madre, prendeva Genmaro sulle ginocchia, gli chiedeva: la vedi com'è bella tua mamma, le vuoi bene?					
n275	Paskvale bi, na sam pomen majke, uzimao Denara u krilo, propitivao ga je: vidiš li kakvu lepu majku imaš, voliš li je?					
n1994	E il vecchio hadži-Zuko, che è già andato due volte alla Mecca e ha oltrepassato i novant'anni, dice che, tempo una generazione, e la frontiera turca arretrerà fino al Mar Nero , quindici giorni di cammino da qui ."					
n1994	A stari Hadži-Zuko , koji je dva puta išao na čabu i kome je prešlo devedeset godina, kaže da neće proći jedan ljudski vijek a turska granica će otići čak tamo na karadenjiz, na petnaest konaka odavle.					
n2029	Questo Nail-bey di Nezuqe, unico maschio dell'anziano bey, fu tra i primi a posare gli occhi su Fatima di Velji Lug.					
n2029	Taj Nailbeg iz Nezuqa, begovski jedinac, bacio je među prvima oko na Fatimu iz Veljeg Luga.					

Figure 5: Display of bilingual text processing using the NER service.

the API-s are opened to the web, they can also be integrated into other applications (e.g. for Python applications by using the requests module). Also, since the complete application is available in open access, other instances of it can be run on user-computers locally (which requires certain packages to be preinstalled) or run as another instance (with the same capabilities) on the web. With each of these methods providing access to the services with the same functionality, users can choose the method that

best suits their needs to access the web services. Here's an example how to send requests using the Python requests module to access the web services:

```
# Define the endpoint and parameters
endpoint = "https://ners.jerteh.rs/ner"
params = {"text": "example_text", "language": "en"}
# Send the request
response = requests.post(endpoint, json=params)
# Print the response
print(response.json())
```

In this example, the endpoint is set to the Named Entity Recognition service and the parameters include the text to be processed and the language of the text. The requests module is used to send a POST request with the parameters as JSON. The response is then printed in JSON format. This example can be adapted to use other services and parameters as needed.

```
import requests
# choose language - lang
# @param ['ca', 'zh', 'hr', 'da', 'nl', 'en', 'fi', 'fr', 'de', 'el', 'it',
'ja', 'ko', 'lt', 'mk', 'nb', 'pl', 'pt', 'ro', 'ru', 'es', 'sv', 'uk', 'sr']
lang = "it"
# choose service option - feat
# @param ['ner', 'nel', 'ner+nel', 'geo']
feat = "nel"
# use api
API_KEY = ["file", "data", "lng", "feat"]
url = 'https://ners.jerteh.rs/api'
params = dict(key=API_KEY, data=data, lng=lang, feat=feat)
res = requests.get(url, params=params)
```

All of the mentioned services offer two different formats for displaying the processing results: HTML and XML. This is demonstrated in the following example. Figure 6 shows the processing of text entered directly into the text box for Italian. A selection of language and the NER service option were also selected, and as a result, HTML was generated. The frame with the resulting HTML contains javascript-powered button that, when clicked, downloads the mentioned XML result to a local computer. Additionally, using the NEL and NER+NEL services, the output is similar, but it includes links to annotated wikidata items. Furthermore, this service also provides a description of the entity by mouse-over event on the entity, by using the description of the corresponding item in wikidata. Current implementation is using description of tagged named entities in English, provided by embedded library, regardless of the text language. Descriptions in English are the most common, because wikidata is the most developed for that language, so it is implemented in this version. In the next versions of the service, the approach will be modified so that the description language corresponds to the text's language.

As with the previous web services, geoparsing is available for both bilingual and monolingual resources. Only recognized named entities of the *LOC* class by *NER+NEL* are displayed on the (HTML-based) map. Figure 7 illustrates geoparsing for a small monolingual text in Italian. It shows the location of the named entities recognized in the text on a map, providing a visual representation for each location mentioned in the text, which can be useful for various research and educational purposes.

It should be noted that for different languages, there may be variations in the recognition of named entities and differences in geoparsing due to several reasons:

- For the given language, there may not be a corresponding item (headword) in the knowledge base for the annotated entity.
- The named entity in Serbian (or any other language with rich inflections) may not be recognized because the system does not recognize inflected forms for the language (such as cases different from the nominative singular: *Srbije*, *Beogradu*, etc.).
- The translation equivalents (in this case study Serbian and Italian) may not be literal, so the named entity may not appear in one of the equivalents (see segment number 1994 in Figure 5 and entity *Mar Nero*).

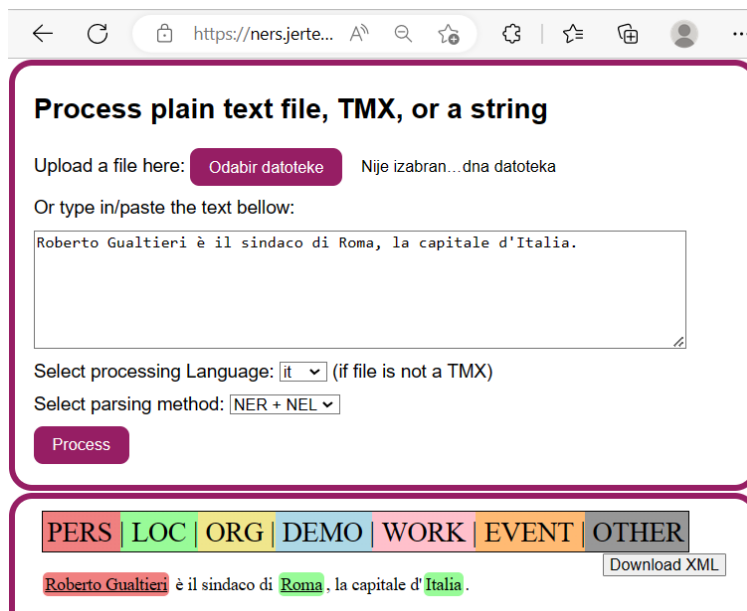


Figure 6: Display of directly entered text processing.

These variations and differences occur due to the specific characteristics of the languages and the resources used. However, the services were developed to handle these variations and differences as much as possible and to provide accurate and useful results for the end users.

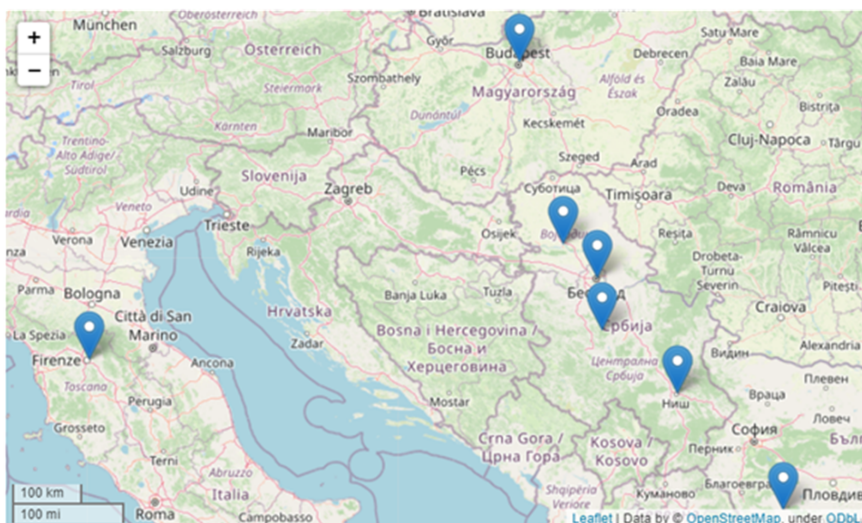


Figure 7: Geoparsing of text and presentation of locations on OpenStreetMap.

5 CQL Corpus Search

Sketch Engine¹⁸ is a widely used tool for exploring the workings of language, based on the analysis of corpora compiled from authentic texts of billions of words. The Sketch Engine allows users to search for a word, phrase or pattern, and results can be presented in various forms such as word sketches, concordances, word lists, frequency graphs, sketch differences, etc. This tool is widely used by researchers,

¹⁸<https://www.sketchengine.eu/>

educators and linguists to analyze and understand language usage, patterns and trends. It was developed by Adam Kilgarriff and his team at the Institute of Formal and Applied Linguistics at the Charles University in Prague and it is a popular tool in the field of corpus linguistics. With its sophisticated features, it allows users to easily extract information from large corpora and analyze it in different ways. (Kilgarriff et al., 2004; Kilgarriff et al., 2014)

NoSketch Engine is an open source edition of the Sketch Engine, which offers core corpus processing and search features, but does not include advanced features such as word sketches and preinstalled corpora. A NoSketch Engine node¹⁹ is installed and maintained by the Society for Language Resources and Technologies JeRTeh, and provides access to several monolingual and bilingual corpora, some of which are available to authorized users only.

The *ItSrNER* corpus in NoSketch is part of speech annotated and lemmatized using TreeTagger²⁰ (Schmid, 1999). The Italian part of corpus is tagged using TreeTagger parameter file prepared by Prof. Achim Stein, University of Stuttgart (Schmid et al., 2007) with 38 tabs in the tagset²¹. The Serbian parametric language parameter file is trained on the harmonized resources, which have been manually annotated within different projects (Stanković et al., 2020), consulting the system of morphological electronic dictionaries of the Serbian language (Krstev, 2008; Vitas and Krstev, 2012). The POS tagset for Serbian part is Universal Dependencies Tagset²²

The ItSrNER corpus can be freely accessed and searched using CQL (Corpus Query Language). Figure 8 presents parallel concordances of ItSrNER corpus for query: *family* (Italian: *famiglia*, Serbian: *porodica*) with an option that NER tags are visible.

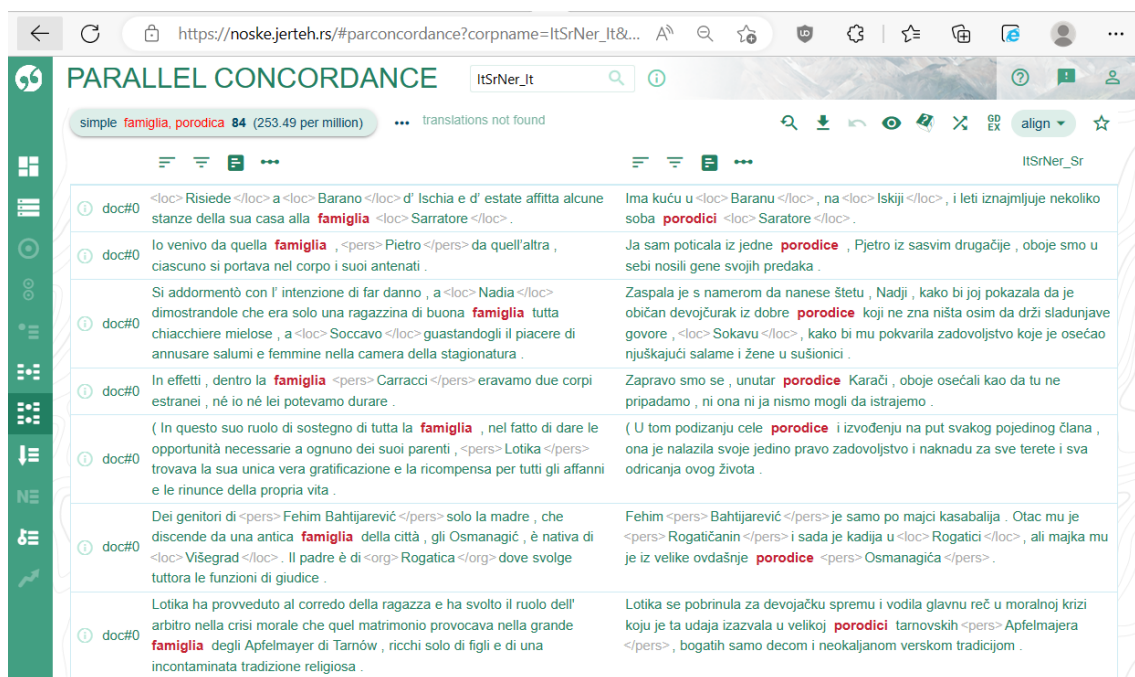


Figure 8: Annotated aligned ItSrNER corpus with NER tags on NoSketch engine platform.

When researching the translation of location names, including those with adjectives, advanced CQL queries can be used to limit the search to the context of annotated location named entities. Figure 9 presents a page with a simple CQL query `[tag="ADJ"]+[tag="NOM"] within <loc/>`, which retrieves concordances with the names of locations in the following form: noun preceded by one or more adjectives. The accuracy of the labeling of named entities, especially the class association of the named

¹⁹<https://noske.jerteh.rs/#dashboard?corpname=srpELTeC>

²⁰<https://www.cis.lmu.de/schmid/tools/TreeTagger/>

²¹<https://www.cis.lmu.de/schmid/tools/TreeTagger/data/italian-tagset.txt>

²²<https://universaldependencies.org/u/pos/index.html>

entity, is not always correct, as can be seen in the first example.

doc#2	" Rum per il Guercio ! " sbraita <loc> Santo Papo </loc> con la voce rauca e il suo accento spagnolo , pensando di essere all' osteria e allargando le mani come se lo crocefiggessero .	Rum za <pers> Ćorkana </pers> ! - derao se <pers> Santo </pers> Papo promuklim glasom , sa španskim izgovorom , misleći da je u mehani i šireći ruke kao da ga razapinju .
doc#2	La moglie poco dopo morì e il fratello pazzo finì nel monastero del <loc> Santo Padre Prohor </loc> .	Naskoro žena umrla , brat u manastiru , <work n="25"> Svetom Ocu </work> Prohoru , umobolan svršio .
doc#3	Mi trovavo ai cancelli del <loc> Nuovo Pignone </loc> , scoppiarono tafferugli , scappai .	Nalazila sam se u blizini kapije Nove železare kada izbiše neredi , ja pobegoh .

Figure 9: Advanced CQL query within <loc> tag.

6 Conclusion

In the paper, we discussed the outcomes of the It-Sr-NER project, which is a web service for annotating named entities for 24 languages and displaying them on a map, with the case study on Italian and Serbian parallel texts. The project was supported by the Common Language Resources and Technology Infrastructure, CLARIN ERIC, and involved a collaboration between the University of Turin and the Society for language resources and technologies JeRTeh. The goal of the project was to improve the teaching of Italian and Serbian languages and to support translation studies. The lack of specific language technologies for the Serbian language has for years been an obstacle in the introduction of the new methodologies in teaching like corpus-based and Data Driven learning. Isolated efforts to incorporate corpora into teaching, although efficient, do not provide enough incentive for researchers and educators in the field of teaching Serbian as a foreign language. At the same time if the students are introduced to corpora and other linguistic tools through proper training, they may gradually develop researcher attitude which allow them to be more creative and to participate actively in the construction of their own learning process.

The primary outcome of the project was the release of a suite of web services for monolingual and bilingual parallel texts available on the CLARIN platform Language Resource Switchboard. Additionally, the project accomplished several secondary objectives that were equally important, such as the creation and publication of a parallel Italian-Serbian corpus, and the development of a web application and service on the JeRTeh platform for language resources and technologies. In total, eight services were created, four for monolingual and four for bilingual resources. These services can process text through direct input at the sentence level, or by processing user-uploaded files. The services also include linking of named entities with wikidata and geoparsing. While the project focused on Serbian and Italian resources, the developed services are capable of processing texts in 24 languages.

Additional research will be conducted to promote the use of the web services and integrate them into teaching. A key objective is to expand the corpus and enhance the model for annotating named entities and linking them to knowledge bases.

Acknowledgements

The authors are thankful to CLARIN ERIC, Common Language Resources and Technology Infrastructure, for supporting our project within the "Bridging Gaps Call 2022". The authors are also grateful to prof. Cvetana Krstev, prof. Duško Vitas, prof. Saša Moderc and Nikola Janković for providing the parallelization.

References

- [de Jong et al.2022] Franciska de Jong, Dieter Van Uytvanck, Francesca Frontini, Antal van den Bosch, Darja Fišer, and Andreas Witt. 2022. Language matters. the european research infrastructure clarin, today and tomorrow. In *CLARIN. The infrastructure for language resources*, pages 31–57. de Gruyter.

- [Delpuch2019] Antonin Delpuch. 2019. Opentapioca: Lightweight entity linking for wikidata. *CoRR*, abs/1904.09131.
- [Draxler et al.2022] Christoph Draxler, Alexander Geyken, Erhard Hinrichs, Annette Klosa-Kückelhaus, Elke Teich, and Thorsten Trippel. 2022. How to connect language resources, infrastructures, and communities. *CLARIN: The Infrastructure for Language Resources*, 1:275.
- [Granger2018] Sylviane Granger. 2018. Has lexicography reaped the full benefit of the (learner) corpus revolution? In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, page 208.
- [Kilgarriff et al.2004] Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. Itri-04-08 the sketch engine. *Information Technology*, 105(116).
- [Kilgarriff et al.2014] Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1(1):7–36.
- [Krstev and Vitas2011] Cvetana Krstev and Duško Vitas. 2011. An aligned english-serbian corpus. *ELLSIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality)*, 1:495–508.
- [Krstev2008] Cvetana Krstev. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries*. Faculty of Philology of the University of Belgrade.
- [Moderc2015a] Saša Moderc. 2015a. Su un modo di tradurre l’avverbio serbo “inače” in italiano: il caso dell’equivalente “altrimenti”. *Università di Belgrado. In Italica Belgradensia*, 1:61–79.
- [Moderc2015b] Saša G. Moderc. 2015b. Elektronski korpus srpskih književnih dela i njihovih prevoda na italijanski jezik. *Anali Filološkog fakulteta*, 27(2):301–316. 15.
- [Nothman et al.2013] Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.
- [Obradović et al.2008] Ivan Obradović, Ranka Stanković, and Miloš Utvić. 2008. Integrisano okruženje za pripremu paralelizovanog korpusa. *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen*, pages 563–578.
- [Perišić Arsić2018] Olja Perišić Arsić. 2018. L’uso dei corpora nella didattica della traduzione: l’esempio del verbo serbo “prijati” e i suoi traduttori italiani. *Italica Belgradensia*, 2018(1):49–64. 3.
- [Perišić2021] Olja Perišić. 2021. Corpora in the classroom-the case of the serbian language for italian speakers. *New Trends in Slavic Studies*, pages 126–137.
- [Perišić2023] Olja Perišić. 2023. *Il corpus per imparare il serbo. Il futuro dell’apprendimento linguistico*. Edizioni dell’Orso.
- [Šandrih Todorović et al.2021] Branislava Šandrih Todorović, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. 2021. Serbian ner&beyond: The archaic and the modern intertwined. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1252–1260.
- [Schmid et al.2007] Helmut Schmid, Marco Baroni, Erika Zanchetta, and Achim Stein. 2007. Il sistema ‘tree-tagger arricchito’-the enriched treetagger system. *IA Contributi Scientifici*, 4(2):22–23.
- [Schmid1999] Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to german. In *Natural language processing using very large corpora*, pages 13–25. Springer.
- [Sinclair1991] J. McH. Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- [Stanković et al.2020] Ranka Stanković, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for serbian. In *Proc. of The 12th LREC*, pages 3947–3955, Marseille, France. European Language Resources Association.
- [Stanković et al.2021] Ranka Stanković, Cvetana Krstev, Branislava Šandrih Todorović, and Mihailo Škorić. 2021. Annotation of the serbian eltec collection. *Infotheca*, 21(2):43–59. 3.
- [Vitas and Krstev2012] Duško Vitas and Cvetana Krstev. 2012. Processing of Corpora of Serbian Using Electronic Dictionaries. *Prace Filologiczne*, XVIII:279–292.
- [Vitas and Lažetić-Pavlović2008] Duško Vitas and Gordana Lažetić-Pavlović. 2008. Resources and Methods for Named Entity Recognition in Serbian. *Infotheca*, 9(1–2):35a–42a, May.
- [Vitas and Poletanović2020] Milica Vitaz and Milica Poletanović. 2020. Data-driven learning the serbian case. *EL.LE*, pages 409–422, april.

Lemmatizing and POS-tagging Akkadian with BabyLemmatizer and Dictionary-Based Post-Correction

Aleksi Sahala

University of Helsinki, Finland
aleksi.sahala@helsinki.fi

Tero Alstola

University of Helsinki, Finland
tero.alstola@helsinki.fi

Jonathan Valk

University of Helsinki, Finland
jonathan.valk@helsinki.fi

Krister Lindén

University of Helsinki, Finland
krister.linden@helsinki.fi

Abstract

We present BabyLemmatizer, a hybrid lemmatizer and POS-tagger for Akkadian, the language of the ancient Assyrians and Babylonians, documented from 2350 BCE to 100 CE. In our approach the text is first POS-tagged and lemmatized with TurkuNLP trained with human-verified labels, and then post-corrected with dictionary-based methods to improve the lemmatization quality. The post-correction also assigns labels with confidence scores to flag the most suspicious lemmatizations for manual validation. We demonstrate that the presented tool achieves a Lemma+POS labeling accuracy of 94%, and a lemmatization accuracy of 95% in a held-out test set. We also apply the lemmatizer to a previously unlemmatized text corpus to test it in practice.

1 Introduction

Application of computational methods to historical text corpora provides interesting opportunities for studying large-scale phenomena that are difficult to perceive through close reading of texts. This often requires careful normalization of the language, because in many past societies spelling conventions were not fully standardized, and the corpora can contain documents written in several synchronic and diachronic variants of the language. The languages can also be morphologically complex, which further complicates even such fundamental tasks as searching for all attestations of a word in the corpus.

One way to normalize historical languages is lemmatization, which labels words with their dictionary forms regardless of their morphology and spelling. In this paper, we present a lemmatizer for Akkadian, an extinct language that was widely used as a lingua franca in ancient Mesopotamia.

The motivation for this tool emerges from close co-operation between the FIN-CLARIN coordinated Language Bank of Finland and the Centre of Excellence in Near Eastern Empires, a University of Helsinki-based research project focusing on the study of the Near East in the first millennium BCE. As part of this co-operation, the Language Bank of Finland collects corpora of ancient Mesopotamian texts written in the Akkadian language in the Korp concordance service.¹ Korp offers several useful functionalities for historians, from flexible search options to generation of statistics from text metadata as well as map views and timelines (Borin et al., 2012).

At present, Korp hosts a version of the Open Richly Annotated Cuneiform Corpus (Oracc),² which comes with human-verified lemmatization. The next Akkadian corpus to be included in Korp is Achemenet,³ which has not been manually lemmatized. The only Akkadian lemmatizer currently available (Tinney, 2019) requires extensive human supervision. To minimize the need for human intervention, our aim is to lemmatize the Achemenet corpus by first training the TurkuNLP's (Kanerva et al., 2021) universal lemmatizer using the available Oracc data, and then applying simple dictionary-based post-correction scripts.

Sahala developed BabyLemmatizer and wrote sections 1, 3, 4, 5, and 6. Valk and Alstola corrected the lemmatizations in manual evaluation. Valk wrote section 2 and Alstola section 2.1. Lindén was the team leader. The current version of BabyLemmatizer is available at <https://github.com/asahala/BabyLemmatizer>.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

¹<https://korp.csc.fi/>

²<http://oracc.org/>

³<http://www.achemenet.com/>

2 The Akkadian Language

Akkadian is an extinct East Semitic language (Hasselbach-Andee, 2021). It is attested in hundreds of thousands of inscriptions primarily from modern Iraq, but also from other sites across the Middle East. The earliest evidence of Akkadian comes in the form of personal names in texts from the Early Dynastic III Period (ca. 2600–2450 BCE). The oldest exemplars of continuous Akkadian text hail from the Sargonic Period (2350–2170 BCE), when the language was adopted for official purposes by the Akkadian Empire. After this period, the language is generally attested in one of its two main dialects: Assyrian (2000–600 BCE) and Babylonian (2100 BCE–100 CE) (Kouwenberg, 2012), both of which can be divided into different stages of development. From the second millennium BCE onward, there is also a literary dialect of Akkadian known as Standard Babylonian (Hess, 2020). Although most historical speakers of Akkadian appear to have lived in modern Iraq, the language served as a scholarly and diplomatic lingua franca in the Middle East for much of the second millennium BCE (Vita, 2020). Vernacular Akkadian died out in the first millennium BCE. Nevertheless, Akkadian continued to be used as a language of scholarship into the first centuries of the Common Era (Geller, 1997).

The corpus of Akkadian texts is vast, numbering approximately 10 million published words (Streck, 2010). This number will only grow as more Akkadian texts in museum collections are published and others are recovered from the Middle Eastern soil. Making this corpus available for computational analysis offers tremendous opportunities for future research. Yet Akkadian presents serious difficulties for automatic reading. Like other Semitic languages, Akkadian morphology employs nonconcatenative root-pattern morphotactics in stem formation and concatenative morphotactics in the attachment of various grammatical affixes to the stems. For example, the verbal form *ludlul* "let me praise (it)!" consists of the first person singular precativ suffix {lu} attached to the preterite stem {dlul}, which is formed from the root *dll* of the verb *dalālu* "to praise". Although the morpheme boundaries are transparent in this example, various morphophonological processes often obscure the underlying structure of the word, complicating recognition of the root radicals (von Soden, 1995). These difficulties are apparent in the following derived surface forms of the verb *warū* "to go up": *umda* "ir" "he commanded", *umtēr* "I assign", *īrama* "he proceeded to" (von Soden, 1995).

Another layer of complexity emerges from the use of the cuneiform script to write Akkadian (Streck, 2021). The cuneiform writing system first developed toward the end of the fourth millennium BCE to represent the Sumerian language, which is unrelated to Akkadian. It was only in the 24th century BCE that cuneiform was adapted to represent Akkadian. The cuneiform script is logo-syllabic. Signs usually represent either a syllable or a logogram. But signs can also represent determinatives, grammatical markers, and phonetic complements. Determinatives mark words as belonging to categories that include male and female personal names, divine names, geographical names, the material of an object, and types of animals. Grammatical markers are attached to logograms and convey information like whether a noun is plural. Phonetic complements can be appended to logograms to suggest to the reader the intended grammatical form of the verb represented by a logogram.

In Akkadian transliteration, logograms are represented in capital letters and named after their base reading values in Sumerian rather than Akkadian. For this reason, the character level relationship between the graphemic and phonemic forms of logographic spellings is typically suppletive. Many logograms are also ambiguous and can have different readings in different contexts. For example, the Sumerian logogram IGI (depicting an eye) can indicate any form of the words *īnu* "eye", *pānu* "front", *māru* "before", and *amāru* "to see". Because Sumerian does not have the same phonemic repertoire as Akkadian, the cuneiform script with its inherited Sumerian values does not align perfectly with the needs of the Akkadian language. The syllabic values of cuneiform signs often collapse phonemically distinct Akkadian consonants like the dentals /t/, /t̄/, and /d/, and the velars /g/, /k/, and /q/. The cuneiform sign with the syllabic value /ig/ can, for example, also represent the syllables /ik/ and /iq/, while the sign /ud/ can also represent /ut/ and /uṭ/. Many cuneiform signs have multiple possible syllabic, logographic, and other values. The correct reading for any sign can only be determined contextually.

Akkadian has few fixed spelling conventions. The Akkadian verbal form *iddin* "(s)he gave it" can, for instance, be spelled syllabically as *id-din*, *i-din*, *id-di-in*, or *i-di-in*. Although Akkadian is generally

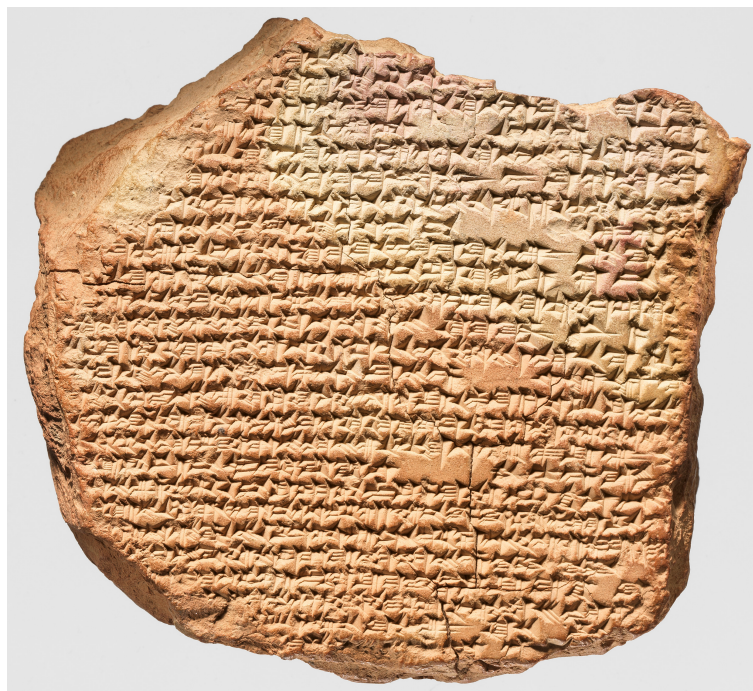


Figure 1: A bilingual (Sumerian and Akkadian) clay tablet from the Neo-Babylonian period written in the cuneiform script (9 x 9.8 x 2.9 cm). The Metropolitan Museum of Art 86.11.313.

written syllabically, scribes sometimes favored the use of Sumerian logograms, especially in certain genres of text. The verbal form *iddin* "(s)he gave it" can therefore also be rendered with logographic and logo-syllabic spellings like SUM and SUM-*in*. The multiplicity of spellings and sign readings in Akkadian pose special challenges to lemmatization and morphological analysis. The training data for any lemmatizer must enumerate the full range of possible logographic, syllabic, and other sign readings, as well as account for the breadth of different spellings.

2.1 Digital Resources

For an extinct language, Akkadian is fairly well resourced, and texts comprising about 3–4 million tokens (words) in total have been digitized.⁴ However, only a fraction of all Akkadian texts exist in a digital format, and even fewer texts have been lemmatized. According to an estimate by Streck (2010), the known Akkadian texts contain up to 10 million words, which indicates that the current digital corpora represent only about one third of the total word mass. Some larger text corpora are Archibab⁵ with 22,500 Akkadian texts, the Cuneiform Digital Library Initiative (CDLI)⁶ with 14,000 texts, Oracc with 13,000 texts, and Achemenet with 5,000 texts.⁷

These four projects highlight the current diversity in the standards of digitizing Akkadian texts. The texts in Oracc have been linguistically annotated and can be downloaded as JSON files, and they are thus well suited for computational analysis without much further processing. The Akkadian texts in Oracc comprise about two million lemmatized tokens in total. At the other end of the spectrum, the Achemenet texts are provided as transliterations with some metadata and occasional translations, and they can only

⁴This figure is our estimate. There are no surveys that report accurate estimates, nor studies that indicate how much overlap different resources have.

⁵<https://www.archibab.fr/>

⁶<https://cdli.mpiwg-berlin.mpg.de/>

⁷There is overlap between the corpora, but the number of duplicates is difficult to estimate. A complete but somewhat outdated survey of Akkadian digital resources is given in Charpin (2014). The number of texts in Oracc was counted in February 2023, and the number of texts in Archibab, CDLI, and Achemenet was retrieved from their websites in January 2023.

be accessed as HTML files published on the website. Many texts in Archibab have been lemmatized, but it is not possible to download the annotated data. The texts in CDLI are only transliterated, but they can be easily downloaded. CDLI is also the largest database of cuneiform language metadata, containing information about 350,000 cuneiform texts or their fragments, including 88,000 written in the Akkadian language.

The limited availability and uneven geographic and chronological distribution of lemmatized texts pose problems for the computational study of the Akkadian language. The majority of texts in Oracc have been written in the Neo-Assyrian period (934–612 BCE), whereas CDLI focuses on texts from the Old-Babylonian (2003–1595 BCE) and Neo-Assyrian periods, Archibab on texts from the Old-Babylonian period, and Achemenet on texts from the Persian period (539–331 BCE). Because a large quantity of lemmatized texts is readily available only from the Neo-Assyrian period, it is currently not possible to do diachronic or cross-cultural studies of the Akkadian language with computational methods. Using the lemmatizer described in this paper, we aim to alleviate this problem by creating a large corpus of lemmatized texts from the Neo-Babylonian (in our context, 626–539 BCE) and Persian periods. For this purpose, we have acquired a corpus of 3,000 texts from Achemenet and are in the process of collecting additional transliterated text corpora from our colleagues.

3 Previous Work

Due to the previously discussed complexity of the Akkadian morphology and script, lemmatization is considered a mandatory step in making any digital corpus of Akkadian searchable or suitable for computational analysis (Maiocchi, 2019). To date, however, only Oracc provides an extensive and downloadable dataset of lemmatized Akkadian texts, totalling about two million lemmatized words. Oracc is lemmatized using a dictionary-based tool known as L2 (Tinney, 2019), which populates new texts with lemmata and part of speech (POS) tags based on a labeled glossary extracted from previously lemmatized texts. Texts are then checked manually word-by-word, filling in lemmata for out-of-vocabulary words and resolving possible ambiguities.

In addition to L2, there are also other tools for Akkadian lemmatization (Sahala, 2021). The earliest proto-type was a two-level morphology by Laura Kataja and Kimmo Koskeniemi, which was also the first attempt to morphologically analyze and lemmatize Semitic discontinuative morphology (Kataja and Koskeniemi, 1988). A more practical dictionary-based lemmatizer was developed by Simo Parpola and Robert M. Whiting for the use of the Neo-Assyrian Text Corpus Project in the late 1980s, but the source code or description of the system has not been published.⁸

A recent morphology-based lemmatizer is Bamman’s finite-state morphology for Old Assyrian (Bamman, 2012), capable of lemmatizing and morphologically analyzing Old Assyrian letters from transliteration. The only finite-state morphology for Babylonian is BabyFST (Sahala et al., 2020). However, at its current state, it is not capable of disambiguating the morphological analyses or lemmatization, and it requires the input text to be normalized into phonological transcription (e.g. *inaddin* “he gives”) instead of transliteration (e.g. SUM or *i-na-din*, *i-na-di-in*), which is the standard way of rendering Akkadian texts in the Latin alphabet.

The system presented in this paper aims to improve four aspects of the existing Akkadian lemmatizers:

1. Handling out-of-vocabulary (OOV) words that are problematic for dictionary-based methods.
2. Handling ambiguity by producing exactly one analysis per word form.
3. Providing a simple self-evaluation of the lemmatization results to minimize the need for manual processing.
4. Offering a possibility to train a specific model for the required task, such as annotation of a certain domain of texts that belong to a specific time period, genre, or dialect.

⁸Personal communication with Parpola.

3.1 Relevance to CLARIN

The Language Bank of Finland aims to improve the accessibility and usability of cuneiform corpora by hosting them in Korp. Currently Korp hosts a version of the Oracc corpora⁹ providing some additional means to work with the cuneiform texts, such as statistics based on text metadata, plots of the provenance of texts on a map (see Figure 2), and search by lemma and POS-tagging by preprocessing previously unannotated texts with the BabyLemmatizer. Currently, Achemenet, a linguistically unannotated corpus of 3,000 Neo- and Late Babylonian texts (from the mid-first millennium BCE), is being added to Korp.

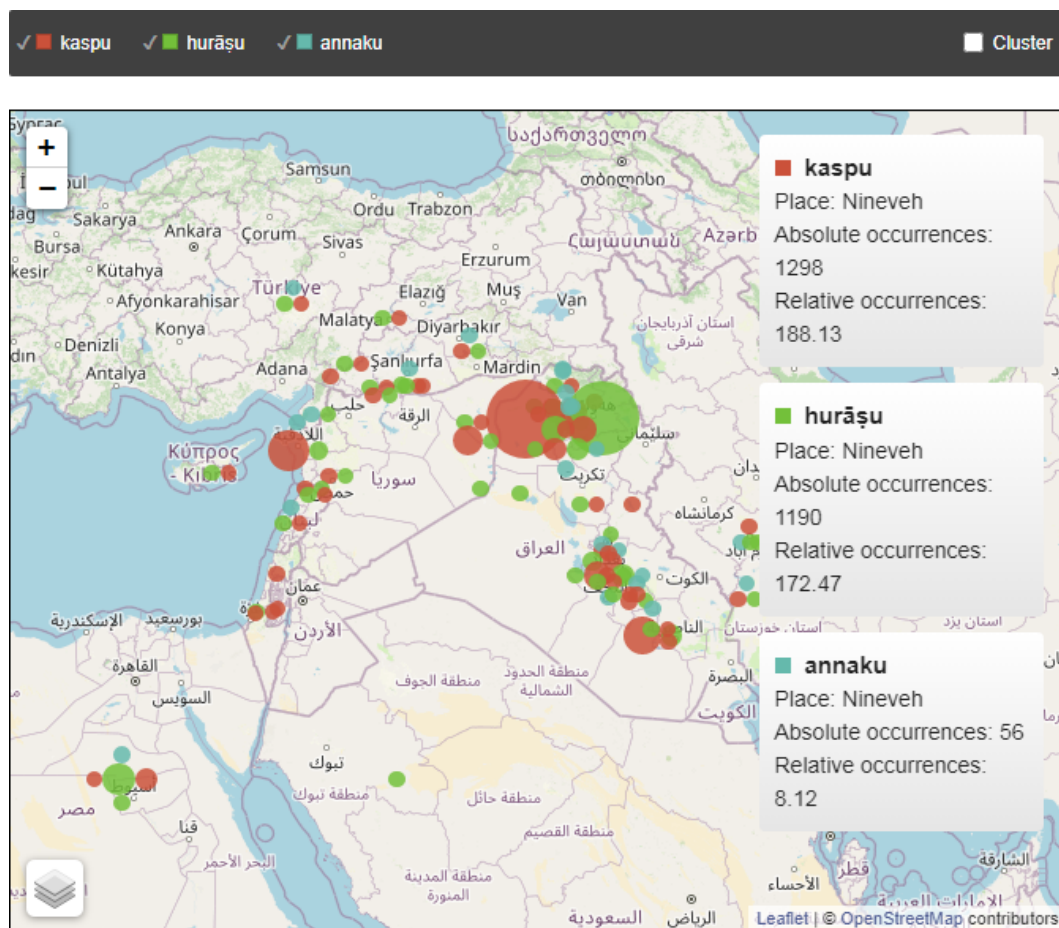


Figure 2: A screenshot of Korp’s map feature showing provenances of texts where Akkadian words *kaspu* ”silver”, *hurāšu* ”gold” and *annaku* ”tin” are mentioned in Oracc.

4 Description of BabyLemmatizer

BabyLemmatizer¹⁰ is a hybrid lemmatizer that utilizes predictive features of neural networks to handle out-of-vocabulary words and dictionary-based lemmatization for previously known word forms. The neural lemmatization is based on a character level representation of the Akkadian transliteration, whereas the dictionary-based method is based on word form tokens. During the lemmatization process, the input stream follows standard Oracc transliteration guidelines¹¹ with the exception of converting determinatives into capital letters as if they were logograms. This conversion aims to reinforce the logographic nature of determinatives and to guide the neural network to not confuse them with similar signs that represent phonetic sequences. For example, capitalizing the determinative of geographical locations {ki}

⁹<http://urn.fi/urn:nbn:fi:lb-2022031705>

¹⁰<https://github.com/asahala/BabyLemmatizer>

¹¹<http://oracc.museum.upenn.edu/doc/help/languages/akkadian/akkadianstyleSheet/>

into {KI}, associates it more closely with the logogram KI "land" than the syllabic reading *ki* of this same sign.

Currently the backbone of our tool is the Turku Neural Parser Pipeline (TurkuNLP) (Kanerva et al., 2018), a state-of-the-art neural lemmatizer built around Dozat's POS-tagger and parser (Dozat et al., 2017). We first train a model for TurkuNLP using Oracc data to provide input text with raw lemmatization and POS-tagging, and then apply this model on the source text and run dictionary-based post-corrections on the result to improve the lemmatization accuracy. In our system, the post-correction involves three distinct steps:

The first step overrides all predictions for in-vocabulary words to minimize the effect of mislearned character level relationships between spellings and their lemmata. We calculate the degree of ambiguity for all lemmatizations in the training data and create a *master glossary* of word forms that have a low degree of ambiguity, and use this dictionary to override all lemmatizations of in-vocabulary words. The degree of ambiguity for a word form is considered to be low, if any lemma+POS label constitutes over 60% of all the labels assigned to it in the training data. At this step, we leave ambiguous words untouched.

The second step aims to assign correct lemmata to words that are known to be ambiguous based on the training data, especially those written with logograms. We calculate co-occurrence probabilities for lemmata and their adjacent POS-tags in the training data, and then assign the most likely lemmata for all word forms in the text based on their POS contexts. We rely on POS-tags instead of surrounding lemmata due to the high POS-tagging accuracy of the TurkuNLP's tagger (ca. 97% for Akkadian), and because the Akkadian corpus is fairly small, which makes using adjacent word-forms or lemmata infeasible. In addition to determining the most likely lemmatization for an ambiguous word form, this step also allows us to reconfirm that our close-to-unambiguous lemmata determined in the previous step are probably correct. This information is later used to score the confidence of our lemmatizations, explained in the next subsection.

Finally, we apply various other post-corrections to the data, such as removing the lemmatization from numbers and words that occur in badly damaged sections of the tablet. These parts are easy to detect, because in the Akkadian transliteration unreadable signs are indicated with the symbol *x*, as in *x-x-in-nu*. This is done to make the lemmatizations more consistent with Oracc conventions, which generally leave too badly damaged and thus unrecognizable word forms unlemmatized. We also heuristically detect some obvious lemmatization errors, such as verbs that show impossible or very unlikely dictionary form patterns. An example of this would be cases in which a word has been labeled as a verb in the POS-tagging process, but lemmatized as if it were a noun. This is possible, because many Akkadian nouns and adjectives derive from verbs. Nonetheless, these can only be flagged for human editors, rather than fixed automatically especially for word forms that do not exist in the training data.

4.1 Confidence Scoring

All lemmata are assigned with a confidence score based on the post-correction steps they have passed. This aims to help Assyriologists in finding the most likely incorrect lemmata from the text and maximize the efficiency of manual lemmatization corrections. OOV logograms and logo-syllabic spellings receive the lowest class of 0 as the relationship between logographic spellings and their lemmatizations is generally suppletive. Syllabic spellings of OOV words receive a confidence score of 1, as they are possible to predict but they cannot be verified by the post-correction process. Of the highest confidence classes, the score of 3 is given to all words that have passed the first post-correction step and thus have been considered to be unambiguous or close to unambiguous. This score is raised to 4, if the lemmatization also passes the second post-correction step and is thus verified to exist in a previously seen part-of-speech context. Remaining in-vocabulary words, namely those with high ambiguity that cannot be resolved by their POS context, are given a confidence score of 2.

5 Evaluation

For evaluation, we train ten models for the first millennium BCE Babylonian texts from Oracc comprising ca. 500,000 Akkadian words in total.¹² The texts represent a wide variety of genres, ranging from astronomical diaries and sign lists to royal inscriptions and legal texts. Everyday texts such as legal transactions are written in the Neo-Babylonian (Late Babylonian) dialect of Akkadian, while literature and royal inscriptions are written in Standard Babylonian, an archaic literary variety of the language (Hess, 2020). We use a text-wise 80/10/10 train/dev/test split and estimate the model’s accuracy against two baseline models by using 10-fold cross-validation.¹³ As the Oracc data is divided into several sub-projects that contain texts that belong to similar genres, we do not shuffle the texts before building our data, but instead pick our 80/10/10 splits in order to ensure that all the sets have a somewhat balanced distribution of different text genres.

Our first **baseline** model is a dictionary-based lemmatizer and POS-tagger that labels the word forms in our test set with their most common lemmata and POS-tags seen in the training data. To measure the effect of our post-corrections, we use **TurkuNLP** without any post-correction scripts as the second baseline model. The results are presented in Table 1.

Model	Lemma	POS	Lemma+POS
Baseline	84.42 \pm 0.33	88.83 \pm 0.31	82.71 \pm 0.34
TurkuNLP	86.19 \pm 1.32	97.32 \pm0.10	85.31 \pm 1.31
BabyLemmatizer	94.94 \pm0.17	97.32 \pm0.10	94.03 \pm0.35

Table 1: Average accuracy (%) based on 10-fold cross-validation.

In addition to validating the lemmatization and POS-tagging accuracy, we examined the distribution of the confidence classes in our evaluation set. Table 2 presents lemmatization accuracies in different confidence classes, as well as the proportion of lemmata that are assigned to each confidence class in our evaluation setting.

Confidence score	0	1	2	3	4
Accuracy	30.66%	56.71%	69.57%	96.25%	98.40%
Lemma-%	0.86%	3.87%	0.49%	52.10%	42.67%

Table 2: Confidence score distribution after all post-correction steps.

The above confidence score distribution also reveals BabyLemmatizer’s capability to handle out-of-vocabulary words. As the confidence scores of 0 and 1 are assigned for out-of-vocabulary syllabic spellings and logo-syllabic spellings respectively, we can observe that the system is able to assign correct lemmata for 30.66% of the OOV logo-syllabic spellings and for 56.71% of the OOV syllabic spellings.

5.1 Manual Evaluation

To test our lemmatizer in practice, we apply it to a sub-corpus of Achemenet comprising 107,778 words.¹⁴ These are primarily Babylonian legal and administrative texts from the Persian period. Although this sub-corpus has a different genre and time period distribution than our previous test sets, the texts were not completely out-of-domain, since our training data included 371 legal documents from the Hellenistic period (late first millennium BCE) comprising 107,403 words, and 87 texts from the Persian period comprising 5,893 words. For administrative texts, our training data comprised only 34 texts totaling 1,734 words. We use a model trained with the same Oracc data and train/dev/test split as in our evaluation setting described above, with an added glossary of Akkadian personal names from the

¹²Every model uses the same hyperparameters.

¹³In this experiment we use the default network architectures for training TurkuNLP’s lemmatizer and tagger.

¹⁴The texts were provided to us by the Achemenet project in December 2020.

Prosobab database (Waerzeggers and Groß et al., 2019). We then generate glossaries of the most common words that were assigned with the two lowest confidence classes and manually correct lemmata and POS-tags for word forms in the glossary file that have a frequency of >3 (for class 0) and >5 (for class 1) in the data. For two text groups in Achemenet (CT 55 and Bel-remanni) both of these frequencies were >2 . There were 315 unique corrected word forms, comprising 3.87% of the unique word forms covering 4.77% (5,037) of the 107,778 words in the sub-corpus.

To measure the accuracy of the lemmatizer and the effect of our manual corrections, we randomly select texts from our lemmatization results amounting to ca. 1,000 tokens for manual evaluation. We first evaluate the initial lemmatization without any manual corrections to the glossaries as a baseline. Then we apply our corrections to the lemmatization results in two ways: first, as a part of our *master glossary* of unambiguous lemmata (used in step 1 of post-correction), and second, by adding our manual corrections to the training data for TurkuNLP to see how much the system can learn from the corrections. The training data is added by first lemmatizing the text with a corrected master glossary and then replacing all words with the lowest two confidence scores with underscores to prevent the neural network from learning likely erroneous lemmatizations. The results are shown in Table 3.

	Lemma	POS	Lemma+POS
Baseline	93.0%	94.6%	90.2%
Glossary Override	96.2%	96.0%	93.8%
Retrained NN	96.6%	96.1%	94.5%

Table 3: Improvement in accuracy after corrections.

As can be seen from Table 3, our Lemma+POS labeling accuracy improves 4.3% when manually correcting only 3.87% of the unique word forms. The final results can be considered satisfactory for our current needs, which are to make the corpus searchable in Korp and to use it for lexical analysis.

6 Conclusions

We presented a hybrid lemmatizer and POS-tagger for Akkadian, and demonstrated an increase of ca. 10% in Lemma+POS labeling accuracy compared with our baseline models. We also tested the lemmatizer on a previously unlemmatized Akkadian corpus with a different chronological and genre distribution than our training data. This test demonstrated that the system can reach a Lemma+POS labeling accuracy close to 95% after minor manual corrections.

As our future work, we plan to try a different input format for training the neural networks. Both the POS-tagger and the lemmatizer used in TurkuNLP split input words into sequences of characters, which is relevant for languages using alphabetic writing systems, but not necessarily for languages that use logo-syllabic writing. Our preliminary tests show that by splitting syllabic signs into sequences of characters but preserving logograms as tokens yields slightly more reliable results for both in-vocabulary and out-of-vocabulary words. We have also experimented with providing more context data (immediately adjacent POS-tags) for the lemmatizer already when the neural network is being trained, which also seems to improve the lemmatization accuracy.

We plan to integrate the morphological analyzer BabyFST to BabyLemmatizer, as BabyLemmatizer can be used to disambiguate morphological annotations at least on the lemmatization and POS-tagging level.

We also plan on training BabyLemmatizer for other cuneiform languages such as Sumerian and Urartian, as well as some Akkadian dialects (Assyrian) and sister languages such as Eblaite.

Acknowledgements

We gratefully acknowledge that the research for this article has been funded by FIN-CLARIN and the Academy of Finland (decision numbers 330727, 336673, and 341798). We thank the Open Richly Annotated Cuneiform Corpus (Oracc) for their efforts in making linguistically annotated cuneiform texts available online and the Achemenet project for providing us with their corpus of transliterated Babylonian texts. We thank Tommi Jauhiainen for providing feedback on a draft version of the article.

References

- David Bamman. 2012. *NLP Lab Report: Akkadian-morph-analyzer*, <https://github.com/dbamman/akkadian-morph-analyzer>.
- Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 474–478, Istanbul. ELRA.
- Dominique Charpin. 2014. Ressources Assyriologiques sur Internet. In *Bibliotheca Orientalis* 71., October.
- Timothy Dozat, Peng Qi, and Christopher D Manning. 2017. Stanford’s graph-based neural dependency parser at the conll 2017 shared task. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, pages 20–30.
- Mark Geller. 1997. The First Wedge. In *Zeitschrift für Assyriologie* 87, pages 43–95.
- Rebecca Hasselbach-Andee. 2021. Classification of Akkadian within the Semitic Family. In J. P. Vita, editor, *History of the Akkadian Language*, pages 119–146. Brill.
- Christian W. Hess. 2020. *Standard Babylonian*. Wiley Sons, Ltd.
- Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 133–142, Brussels, Belgium, October. Association for Computational Linguistics.
- Jenna Kanerva, Filip Ginter, and Tapio Salakoski. 2021. Universal lemmatizer: A sequence-to-sequence model for lemmatizing universal dependencies treebanks. *Natural Language Engineering*, 27(5):545–574.
- Laura Kataja and Kimmo Koskenniemi. 1988. Finite-state Description of Semitic Morphology: A Case Study of Ancient Accadian. In *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*.
- Bert Kouwenberg. 2012. Akkadian in General. In Weninger S., editor, *The Semitic Languages: An International Handbook*, pages 330–339. De Gruyter Mouton.
- Massimo Maiocchi. 2019. Thoughts on Ancient Textual Sources in Their Current Digital Embodiments]. In S. Valentini and G. Guarducci, editors, *Between Syria and the Highlands: Studies in Honor of Giorgio Buccellati and Marilyn Kelly-Buccellati*, pages 262–268. CAMNES.
- Aleksi Sahala, Miikka Silfverberg, Antti Arppe, and Krister Lindén. 2020. BabyFST-towards a finite-state based computational model of ancient babylonian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3886–3894.
- Aleksi Sahala. 2021. *Contributions to Computational Assyriology (PhD Thesis)*. University of Helsinki.
- Michael P. Streck. 2010. Großes Fach Altorientalistik: Der Umfang des keilschriftlichen Textkorpus. In *Mitteilungen der Deutschen Orient-Gesellschaft zu Berlin* 142, pages 35–58, page 35–58.
- Michael P. Streck. 2021. Akkadian and Cuneiform. In *History of the Akkadian Language*. In J. P. Vita, editor, *History of the Akkadian Language*, pages 66–74. Brill.
- Steve Tinney. 2019. *L2: How it Works*, <http://oracc.org/doc/help/lemmatizing/howl2works>.
- Juan-Pablo Vita. 2020. *History of the Akkadian Language (2 vols)*. Brill.
- Wolfram von Soden. 1995. *Grundriss der akkadischen Grammatik (3rd edition)*. Pontifical Biblical Institute, Rome.
- Caroline Waerzeggers and Melanie Groß et al. 2019. *Prosobab: Prosopography of Babylonia (c. 620-330 BCE)*, <https://prosobab.leidenuniv.nl>.

Developing Resources for Measuring Text Readability in Sesotho

Johannes Sibeko

Department of Linguistics and Applied Linguistics
Nelson Mandela University
Gqeberha, South Africa
johannes.sibeko@mandela.ac.za

Abstract

This article presents a work-in-progress doctoral project that explores measuring text readability in Sesotho, a Bantu language spoken by more than 10 million speakers across Southern Africa. The main project adopts a classical readability formulas approach to text readability analysis. We aim to adapt nine existing readability metrics into Sesotho using English as a higher-resourced helper language. So far, five resources have been developed as part of the study. The rule-based and the T_EX-based syllabification systems, the syllable annotated wordlist, and the grade 12 exam reading comprehension and summary writing corpus have been published on the South African Centre for Digital Language Resources' (SADiLaR) online repository. The machine-translated corpus is still under development. This article describes the progress of the PhD project by overviewing the basic digital language resources developed for the project. The metrics under consideration for adaptation into Sesotho are also briefly discussed.

1 Introduction

Automated text readability evaluation has been applied in different application domains such as finding educational materials (Collins-Thompson, 2014). The scholarship of text readability has continued for over a century (Collins-Thompson, 2014; De Clercq and Hoste, 2016). To date, more than 200 metrics have been developed (DuBay, 2004). However, indigenous African languages have been neglected in this area (Sibeko and Van Zaanen, 2021). As a result, matching texts of the right level of readability to readers such as books for learning and teaching in languages where no metrics are available depends on each assessor's intuition. Undoubtedly, intuition-based choices are likely to be flawed, inconsistent and influenced by the content of the text. Unfortunately, although textbooks are the most important teaching material in language teaching discourse, inappropriate textbook use can deskilling both language teachers and language learners (Mohammed et al., 2022). Deskilling can be propelled by a mismatch between textbooks and intended readers (Zamanian and Heydari, 2012). Such dissonances can result from incorrect levels of readability which can be expected when texts are chosen based on intuition. To reduce the chances of deskilling the language reader, it is essential to have a system for objectively estimating the readability of reading texts of varied lengths such as textbooks and comprehension texts.

Post-graduate research projects have made significant contributions to the scholarship of text readability. Researchers have used various methods to study text readability, including classical readability metrics (Kondru, 2006; Feng, 2010; Janan, 2011; Bendová, 2021), machine learning (Sjöholm, 2012; Andova, 2017), Natural Language Processing (Dios, 2016), deep learning (Alkaldi, 2022), and eye-tracking (Newbold, 2013). Furthermore, other research has adapted existing readability metrics to lower-resourced languages (Bendová, 2021). Unfortunately, research in lower-resourced languages such as Sesotho is restrained by the lack of training data. As a result, machine learning approaches are not feasible.

This article reports on a work-in-progress doctoral project on measuring text readability in Sesotho using classical readability metrics. A brief overview of writing in Sesotho is presented (Section 2) followed

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

by a brief synopsis of the classical readability metrics that are adapted in the main project (Section 3). Then the relevance of the project to CLARIN via SADiLaR is briefly highlighted (Section 4) followed by a discussion of the resources produced as part of the project (Section 5). The article is concluded with a discussion of future works (Section 6).

2 Contextualising Sesotho

Sesotho is used by more than ten million speakers in a few countries in Southern Africa including South Africa, Lesotho, and Zimbabwe (Marupi and Charamba, 2022; Sibeko and Setaka, 2022). In fact, it is one of the official languages in Lesotho, Zimbabwe, and South Africa. It is used as a language for learning and teaching in these countries as either a mother tongue, second language, or marginalised language. It is also used for media, political, religious, and other uses. Even so, a recent investigation of the Sesotho Basic Language Resource Kit (BLARK) content has revealed that there is a severe shortage of digital language resources available for Sesotho (Sibeko and Setaka, 2022). As such, Sesotho remains a low-resourced language (Roux and Bosch, 2019; Sibeko and Setaka, 2022). Consequently, automating the process of objectively investigating text readability in Sesotho using classical readability metrics requires the development of a few basic language resources.

In addition to the lack of necessary resources which hinders the development of objective automated metrics for measuring text readability in Sesotho, there are two widely recognised orthographies for Sesotho. The two orthographies are differentiated by the two countries with the most speakers of Sesotho. They are therefore labelled accordingly as the South African Sesotho (SAS) orthography and the Lesotho Sesotho (LS) orthography. The main differences between the two orthographies include the use of *w* and *y* in the SAS orthography as opposed to the use of *o* and *e* in the LS orthography for representing semi-vowels. This is exemplified in example 1 below.

- (1) *Ke wena le yena.* - SAS.
It's you and her.
'It's you and her.'

Ke oena le eena. - LS.
It's you and her.
'It's you and her.'

The LS orthography also uses *l* in place of *d*, and *c* in place of the digraph *tj*. The differences are exemplified in example 2 below.

- (2) *O dula a tjha.* - SAS
He always is burning.
'He is always burning.'

O lula a cha. - LS
He always is burning.
'He is always burning.'

Although differences such as preferences for certain single letters do not affect the results of the metrics adapted in the main project, the different representations of the semivowels will affect syllable identification. Furthermore, the use of single letters in one orthography and the use of digraphs in another orthography may affect average word lengths. Beyond orthography, there may be region-based vocabulary variations in Sesotho. For instance, see Lemeko (2018) for a discussion of region-based variations of SAS. Nonetheless, these variations are not within the scope of the current project. The research focus is limited to the effects of orthographic conventions.

3 Classical Readability Metrics

The doctoral project described in this article focuses on how text readability could be measured in Sesotho. An automated process for measuring text readability in Sesotho is desired. We believe that classical readability metrics are a good place to start. We identified a total of nine classic readability formulas for adaptation into Sesotho. These metrics are used in the Python 3.2 readability package¹. All nine metrics have been used in previous research on South African educational texts, for instance, see Sibeko and Van Zaanen (2021). The readability metrics that we hope to include in our web-based platform for measuring Sesotho text readability are briefly described below. In-depth discussions of these metrics are provided elsewhere, for instance, *see* Heydari (2012) and Zamanian and Heydari (2012).

3.1 Syllable-Based Metrics

Four of the readability metrics identified are based on syllable information. The metrics are described below.

3.1.1 Flech-Kincaid Grade Level (FKGL)

The FKGL uses US grades for labelling readability levels (Kincaid et al., 1975). For example, a score of 10 corresponds to the tenth-grade (Toyama et al., 2017). The FKGL metric uses the following formula:

$$\text{FKGL} = 0.39\left(\frac{\#tokens}{\#sentences}\right) + 11.8\left(\frac{\#syllables}{\#tokens}\right) - 15.59$$

The process for calculating readability follows four steps. First, the total number of words is divided by the total number of sentences and multiplied by the weight given to sentence difficulty, i.e., 0.39. Second, the total number of syllables is divided by the total number of words and multiplied by 11.8, which is the weight given to average word difficulty, that is, the average number of syllables per word. In the third step, the resulting numbers from the first and the second steps are added together. Finally, 15.59 is subtracted from the result of step 3 (Boles et al., 2016).

The FKGL metric was developed for the United States Navy. However, it is suitable for use in multiple contexts including educational contexts (Zhang et al., 2019).

3.1.2 Flesch Reading Ease (FRE)

Flesch's (1948) FRE is calculated using the following formula:

$$\text{FRE} = 206.835 - 1.015\left(\frac{\#tokens}{\#sentences}\right) + 84.6\left(\frac{\#syllables}{\#tokens}\right)$$

The FRE formula outputs scores between zero and 100 (Flesch, 1948). While a text with a score of 100 should be easily readable to a language learner with a fourth-grade education, a text with a score of 0 requires at least a college graduate level for reading with ease.

FRE is one of the most used classical readability formulas. In fact, when combined, the FKGL and FRE can be used for both first and second-language texts (Greenfield, 2004). Both FRE and FKGL are integrated into Microsoft Office (Bendová, 2021). As a result, they can be easily used by anyone who uses Microsoft office products. Furthermore, the FRE metric is the most adapted to other languages (Bendová and Cinková, 2021). For instance, it has been adapted to Italian, French, Spanish, German, Russian, Danish, Bangla, Hindi, and Japanese.

3.1.3 Gunning Fog Index (GFI)

The GFI identifies foggy words which are words comprised of more than two syllables (Zhang et al., 2019; Gunning, 1952; Gunning, 1969; Gunning, 2003). The GFI follows four steps. First, the number of words used per sentence is averaged. Second, the number of foggy words is counted. Third, the percentage of foggy words in the sample is calculated. Finally, the totals are added and multiplied by 0.4 (Eleyan et al., 2020). The following equation is used in the GFI:

¹<https://github.com/andreasvc/readability/>

$$\text{GFI} = 0.4\left[\left(\frac{\#tokens}{\#sentences}\right) + 100\left(\frac{\#complex-words}{\#words}\right)\right]$$

The readability score generated by the English equation above typically falls within the range of 6 to 20. A score of 6 indicates that the text is suitable for a sixth-grade reading level, while a score of 20 or higher suggests that the text is appropriate for advanced readers, such as those in university postgraduate programs.

3.1.4 Simple Measure of Gobbledygook (SMOG)

When calculating SMOG for long texts, three samples are used, one from the beginning of the text, one from the middle, and one from the end of the text ((Mc Laughlin, 1969; Zhou et al., 2017). Each sample comprises ten sentences. The samples are used to calculate SMOG using the following formula:

$$\text{SMOG} = 3.1291 + 1.043\sqrt{\#polysyllabicwords * \left(\frac{30}{\#sentences}\right)}$$

Polysyllabic words as indicated in the formula refer to words with more than two syllables (Kasabwala et al., 2012). The SMOG formula also outputs US grade levels.

3.2 Word-Length-Based Metrics

Four of the selected classical readability metrics are based on word lengths. The four metrics are described below.

3.2.1 Lasbarhetsindex (Lix) and Rate Index (Rix)

Lix was originally developed for Swedish (Björnsson, 1983). For access, we use the English version as a point of reference. It is suggested that ten samples comprising ten sentences each be analysed when estimating both Lix and Rix (Anderson, 1983). The Lix and the Rix formulas pay special attention to ‘long words,’ that is, words that have more than six characters. The Lix formula is presented in table 1.

Lix	Rix
$\text{Lix} = \left(\frac{\#words}{\#sentences}\right) + \left[\frac{\#longwords}{\#words} * 100\right]$	$\text{Rix} = \frac{\#longwords}{\#sentences}$

Table 1: The Lix and the Rix formulas

The Lix formula outputs numbers that are then converted to grade levels. Anderson (1983) states that fractions can be ignored. This is particularly important in instances where adjusting the scores changes the predicted grade. For instance, when adjusting a Lix score of 47.99 to 48.0, the predicted grade level changes from the 11th to the 12th grade.

The Rix metric is an adaptation of the Lix metric (Courtis, 1987; Anderson, 1983). It considers the ratio of long words to the number of sampled sentences. While shorter texts may be considered as a whole, longer texts may use sentence sampling methods. The Rix metric assigns grade levels through the formula presented in table 1.

3.2.2 Coleman-Liau Index (CLI)

The CLI metric also uses a sampling method (Coleman and Liau, 1975). First, the text is divided into shorter samples of 100 words each. Second, the samples are counted. Third, the number of characters in each word from the samples is calculated. Fourth, the number of characters per word is divided by the number of samples. Fifth, the number of sentences is counted. Sixth, the number of sentences is divided by the number of samples. Finally, the results are applied to the following formula:

$$\text{CLI} = 0.0588\left(\frac{\#letters}{\#samples}\right) - 0.296\left(\frac{\#sentences}{\#samples}\right) - 15.8$$

According to Coleman and Liau (1975), samples should end with complete sentences. As a result, CLI samples may contain a little less or more than 100 words depending on the last complete sentence sampled.

3.2.3 Automated Readability Index (ARI)

ARI is derived from fractions representing predictions of word and sentence difficulty (Kaur et al., 2018; Smith and Senter, 1967). The process follows a few steps. First, sentence lengths are averaged and multiplied by 0.5. Second, word lengths are averaged and multiplied by 4.7. Third, the totals are combined and 21.43 is deducted. The grade level is assigned through the following formula:

$$\text{ARI} = 4.7\left(\frac{\#letters}{\#words}\right) + 0.5\left(\frac{\#words}{\#sentences}\right) - 21.43$$

Letters as used in CLI and ARI, refer to all letters and numbers that build words (Zhang et al., 2019). Thomas et al. (1975) describe it as strokes representing each word.

3.3 Frequency-List-Based Metric

One frequency-list-based metric was identified from Python 3.2's readability package. The metric is described below.

3.3.1 Dale-Chall Index (DCI)

The DCI metric uses a frequency list (Dale and Chall, 1948). The frequency list is based on a list of 3000 words that a grade 4 learner is expected to be familiar with (Stocker, 1971). Difficult words are considered as those that do not appear in the list. Variations include words in plural forms, verbs that end in -s, -ed, -ing, and -ied, adverbs that end in -ly, names of both people and organisations [note that organisation names are counted only two times per 100-word sample], abbreviations, and compound words (Barry and Stevenson, 1975). DCI is computed using the following formula:

$$\text{DCI} = 0.0496\left(\frac{\#words}{\#sentences}\right) + [11.8\left(\frac{\#difficultwords}{\#words}\right) * 0.1579] + 3.6365$$

It is advisable to use the whole text when texts are too short for sampling. For longer texts, one may sample four sets of 100 words per 2000 words (Barry, 1980; Dale and Chall, 1948).

3.4 Summary

Two important things can be noted from the nine classical readability metrics briefly overviewed in this article. First, the metrics use specific processes for estimating appropriate readability and grade levels. It is important to consider these processes. For instance, this is useful when considering the minimal corpus size necessary for adapting the metrics to Sesotho. Second, the metrics use specific weights that may need to be adapted to Sesotho. For instance, syllable lengths may have minimal effect on the level of readability and therefore need to carry minimal weighting. Additionally, it is not possible to investigate the effect of syllable lengths on Sesotho texts without a system for identifying the syllables. For this reason, we need the resources that are developed in the doctoral project discussed in this article.

Moreover, it is important to consider the expected outputs from the formulas. This is most important for formulas that do not output grade levels, for instance, the FRE, Lix and Rix. For these metrics, we may need to redefine the conversions to suit the context of Sesotho and to reflect South Africa's grade levels. In spite of being used in multiple contexts, the classical readability metrics have been criticised for failing to measure comprehension (Tanprasert and Kauchak, 2021). Furthermore, the use of frequency lists, such as in the DCI list of common words, has been criticised for failing to account for specialised meanings (Yan et al., 2006). Even so, the classical readability metrics remain relevant to our project since our focus is not on meaning or comprehension but on the ease with which the text can be read.

4 Relevance to SADiLaR

This PhD project is conducted at North-West University which hosts the South African Center for Digital Language Resources (SADiLaR). SADiLaR is an observer at the CLARIN European Research Infrastructure Consortium. North-West University functions as a hub of a network of linked nodes for

SADiLaR. There are currently six nodes including four universities and two independent research entities. SADiLaR is a national center supported by the South African Department of Science and Innovation as part of the South African Research Infrastructure Roadmap (Wilken et al., 2018; Roux and Bosch, 2019). It has an enabling function with a focus on all official languages of South Africa (Roux and Ndinga-Koumba-Binza, 2019). It supports research and development in language technologies and language-related studies in the humanities and social sciences. The center impacts three domains, namely, (i) humanities and social sciences, (ii) language technology, and (iii) socio-economic domains. This doctoral project benefits from SADiLaR’s humanities and social sciences domain which focuses on building research capacity. For instance, several capacity-building training opportunities were freely provided by SADiLaR. The project also benefits from the language technology domain which focuses on the development of high-level resources and NLP tools for use in applications. For instance, some of the resources discussed in this article resulted from collaborative work with experts from SADiLaR. Furthermore, the project has contributed to the development of digital language resources as part of the main project in adapting classical readability metrics to Sesotho.

5 Developing Resources for Sesotho

This section describes five resources developed for the project. Two syllabification systems are described, followed by three annotated datasets.

5.1 Syllabification Systems

A survey of Sesotho digital language resources listed on SADiLaR’s repository web interface indicated the absence of syllabification systems for Sesotho² (Sibeko and Setaka, 2022). For this reason, previous assessments of the readability of Sesotho texts using classical readability metrics relied on the manual extraction of textual properties such as syllable information. Sadly, using annotators to manually extract Sesotho syllable information from written texts is laborious (Krige and Reid, 2017). Additionally, reliance on such manual methods for extracting textual properties would not suffice for the envisaged automated tool. For this reason, two syllabification systems were developed. The systems are briefly described below.

As a tonal language, Sesotho carries tone by vowels and nasal consonants (Guma, 1982; Sekere, 2004; Mohasi et al., 2011). According to Guma (1982), nasal consonants, that is, two simple nasal consonants *n* and *m* and two complex nasal consonants *ŋ* and *ɲ*, and the lateral consonant *l* can occur as syllables. Furthermore, vowels [a, e, i, o, u, ɪ, ε, ɔ, and ʊ] can function as syllables (V). The vowel-only syllables can occur at word initial, word medial, and word-final positions. Nasal consonant-only (C) syllables can also occur in these positions. However, only the complex nasal consonant *ŋ* can occur at word-final position (Demuth, 2007). Finally, syllables can be composed of consonants and vowels (CV). Table 2 presents the syllable types, subtypes, and examples for each subtype. The syllable boundaries are indicated by the use of dashes (-).

We based our syllabification rules on Guma’s (1982) syllable types. For testing the system, we extracted syllabification information from Chitja’s (2010) dictionary. The process for extracting syllable information from the dictionary and creating a wordlist is described in section 5.2.1. The wordlist represented all syllabification types presented in Table 2. The rule-based system achieved an accuracy rate of 99.69%. We also experimented with a T_EX-based approach. We used the wordlist [see section 5.2.1] for training and testing the machine learning system. This system achieved an accuracy rate of 78.92%. The lower accuracy rate of the T_EX-based system is attributed to two unavoidable shortcomings. First, we noticed that there was some human oversight while manually cleaning the training corpus. Second, the T_EX-based system cannot handle single-letter syllables at the beginning or end of words. Both systems are publicly available on SADiLaR’s repository (see Sibeko and Van Zaanen (2022a)).

²SADiLaR indexes both publicly available, and privately hosted digital language resources. For resources that are not publicly accessible, only metadata is indexed. The contact details of the host and sometimes the creators are then provided in case one needs access to the resource. The repository can be accessed online at <https://repo.sadilar.org/>.

Type	Sub-types	Input	Syllabified	English
V	word-initial vowel	<i>ama</i>	<i>a-ma</i>	touch
	consecutive vowels	<i>baena</i>	<i>ba-e-na</i>	brethren
	word-final vowel	<i>letswai</i>	<i>le-tswa-i</i>	salt
CV	one consonant - one vowel	<i>panana</i>	<i>pa-na-na</i>	banana
	one consonant - semi-vowel- one vowel	<i>lwana</i>	<i>lwa-na</i>	fight
	two consonants - one vowel	<i>tlala</i>	<i>tl-a-la</i>	hunger
	two consonants - semi-vowel- one vowel	<i>shwang</i>	<i>shwa-ng</i>	dieing
	three consonants - one vowel	<i>tlhase</i>	<i>tlha-se</i>	spark
C	three consonants - semi-vowel- one vowel	<i>tshwela</i>	<i>tshwe-la</i>	spit
	nasal consonant n, m - non-nasal consonant	<i>ntate</i>	<i>n-ta-te</i>	father
	nasal consonant n, m - nasal consonant	<i>mme</i>	<i>m-me</i>	mother
	nasal consonant n - complex nasal consonant	<i>nnyatsa</i>	<i>n-nya-tsa</i>	disrespects me
	complex nasal consonant ŋ - vowel	<i>ngala</i>	<i>nga-la</i>	abandon
	complex nasal consonant ŋ- non-nasal consonant	<i>mangmang</i>	<i>ma-ng-ma-ng</i>	so so
	word-ending complex nasal consonant ŋ	<i>hang</i>	<i>ha-ng</i>	once
	consecutive lateral consonants l	<i>llela</i>	<i>l-le-la</i>	weep for

Table 2: Syllabification rules and examples.

5.2 Annotated Datasets

5.2.1 Syllabified Wordlist

As part of developing the syllabification systems, we developed a gold-standard syllable information annotated corpus. We extracted dictionary entries and syllable information from *Bukantswe ya Machaba ya Sesotho* ‘The international dictionary of Sesotho’ (Chitja, 2010). Each dictionary entry contains valuable pieces of information. See, for instance, example 3 below.

- (3) *Diepollo* (*di-e-pu-l-law*) /exhumation/ *Ketso ya ho epolla kapa ho ntsha ntho e epetsoeng tlasa mobu. Ketso ya ho ntsha bafu mabitleng. (bap. Kepollo).* - LS
‘**Diepollo** (di-ep-ul-law) /exhumation/ The act of unearthing things buried under the soil. Acts of digging up deceased people from their tombs. (**comp.** Kepollo).’

In example 3, the dictionary entry indicates the Sesotho word, *Diepollo* ‘exhumation’, then provides pronunciation information in brackets (*di-e-pu-l-law*), the English translation ‘exhumations’, the definition ‘*Ketso tsa ho epolla kapa ho ntsha ntho tse epetsoeng tlasa mobu. Ketso ya ho ntsha bafu mabitleng.*’ Which translates to ‘[T]he act of unearthing things buried under the soil. Acts of digging up deceased people from their tombs.’ Finally, a similar word is provided, that is ‘*bon. kepollo*’ which in this case is the singular form: ‘exhumation’. For our project, we extracted the dictionary entries, followed by the pronunciation information as in the example below

- (4) *Diepollo* (*di-e-pu-l-law*) - SAS
Exhumations (ex-hu-ma-tions)

As illustrated in example 4, pronunciation information was not always consistent with orthographic conventions. In instances of such inconsistencies, the wordlist was manually cleaned on a word-for-word basis to ensure consistent orthography. For instance, we altered the pronunciation information at example 4 above. That is, we adjusted the third syllable which illustrated a high tone *o* by using the letter ‘u’, as in *pu* and the fifth syllable which indicated a lower tone *o* by using the digraph ‘aw’. The modified syllables are presented in example 5 below:

- (5) *Diepollo* (di-e-po-l-lo) - SAS
'Exhumations'

Some pronunciation information included words ending in non-syllabic consonants, others changed the spelling such as in example 4, and others had incorrectly placed syllable boundaries. All of these issues were manually checked and fixed. After manual cleaning and fixing orthographic inconsistencies, we obtained a total of 13 551 words. The cleaned wordlist was also uploaded onto SADiLaR's repository (see Sibeko and Van Zaanen (2022b)).

5.2.2 Reading Comprehension and Summary Writing Texts

We were granted access to grade twelve exam question papers by the South African National Department of Basic Education (DBE). Grade twelve is the high school exit grade in South Africa. We have since extracted reading comprehension and summary writing texts from the exam question papers. We did this for all eleven official languages of South Africa³. The texts in our corpus are split into two categories, that is, the home language (HL) and the first additional language (FAL). Previous research indicated that the English exam texts show consistently lower readability levels for the HL texts as opposed to the FAL texts (Sibeko, 2021; Sibeko and Van Zaanen, 2021). The lengths of texts in the collection vary according to the orthographies such as disjunctive and conjunctive, and text types such as reading comprehension and summary writing. Consistently, the lengths of summary texts are about a third of the reading comprehension texts in all eleven languages. The corpus has been uploaded to SADiLaR's repository (see Sibeko and Van Zaanen (2022c)). We hope that the differences in text readability and linguistic complexity are uniform throughout the different languages.

5.2.3 Machine-Translated Corpus

Previous studies evaluating the text readability of Sesotho texts, for instance, Krige and Reid (2017) and Reid et al. (2019), assumed that classical readability metrics that are based on syllable information and word-length-based textual properties can be directly used in Sesotho without taking the differences between the superficial textual features of English and Sesotho into consideration. However, it is evident from other studies adapting the weights of these textual features to language-specific conventions that the metrics cannot be applied to new languages without adjustments. The syllabification systems described in this article enable the automatic identification and counting of syllables. This is already different from previous research on Sesotho text readability. However, we still aim to adapt the metrics to the specific context of Sesotho. Such an adaptation is important for accounting for differences in superficial text properties between Sesotho and English. A gold-standard corpus with clear levels of text difficulty is needed to develop an automated readability model (Van Oosten et al., 2010; François and Fairon, 2012). Unfortunately, Sesotho, like other LRLs, does not have corpora readily annotated with levels of difficulty (Filho et al., 2016). The use of translated texts may provide a solution to this lack of levelled texts. For instance, the texts can be easily levelled according to grades in English.

It was observed that when Obonerva's (2006) readability model that was trained on fiction texts was evaluated on non-fiction texts, exaggerated readability levels were observed (Solovyev et al., 2018). Since we hope that education stakeholders such as teachers, parents, learners, textbook authors, and examiners can use our envisaged automated tool for analysing Sesotho text readability, we are training our models on educational texts. We are relying on texts collected as part of Sibeko and Van Zaanen's (2022c) corpus described in section 5.2.2 above. We identified texts from grade 12 Sesotho HL and FAL examinations from the collection. As a rule of thumb, we followed Zamanian and Heydari's (2012) guideline that a text should have at least 200 words for metrics like FRE and FKGL to be applied successfully. As a result, we identified texts of no less than 200 words each. In the end, we could only use longer reading comprehension and summary texts.

For an illustration of the texts collected, table 3 presents the original Sesotho summary writing text from the December 2012 question paper together with the unedited translation from Google translate. In

³Since 1994, South Africa recognises eleven official languages, that is, Afrikaans, English, IsiNdebele, IsiXhosa, IsiZulu, Sepedi, Sesotho, Setswana, Siswati, Tshivenda, and Xitsonga. Efforts are being made to officialise Sign language as the twelfth official language of South Africa.

Ho phedisana le baahisane ba mona	Living with your neighbours
<p><i>Ho bohlokwa ho ba le dikamano tse ntle le baahisane. Le phela mmoho mme le lokela ho thusana nakong tsa mathata. Leha ho le jwalo, dikamano di ba le diphepetso ha moahisane a rata ditaba. Kopana le moahisane kgafetsa le arolelane mehopollo ya ho ntlafatsa maphelo a lona. Ka ho arolelana tlhahisoleseding, moahisane o tla lemoha hore o sebetsa ka thata ho fumana seo o nang le sona. Ebang a batla ho eketsa ntle kapa ho reka ho itseng, mo eletse hore a ka fumana hokae dintho ka theko e tlase. Ka tsela ena o tla bona hore o a mo kgathalla ebile o a mo tshhetsa. Ha le arolelana ditoro le ditabatabelo, nnetefatsa hore dipuo tsa hao ha di mo fe pelaelo ya hore o a ikgantsha, ho seng jwalo o tla batla tsela ya ho o sitisa. Tse ding tsa dintho e be makunutu a hao. Ha a o botsa dipotso ka se itseng, o fane ka karabo e teletsana e arabang dipotso tsa hae. Ha ho na le se sa o kgotsafatseng, bua hantle o ikokobetse, o rarolle qaka ka tsela e ke keng ya baka kgohlano. Ha moahisane a nahana hore o motho wa maemo a hodimo, se etsa dintho tse nnetefatsang mohopollo oo wa hae. Dula o ikokobeditse ka dinako tsohle, o se ke wa iketsa betere ho mo feta. Ha moahisane enwa a ka bua leshano ka wena, iphanyane o emele mohla e mong a o tobang mme o nke mohato wa ho bua ka seo.</i></p>	<p>It is important to have good relations with your neighbors. You live together and should help each other in times of trouble. However, relationships have challenges when the neighbor likes the news. Meet your neighbor often and share ideas to improve your lives. By sharing information, your neighbor will realize that you are working hard to get what you have. Whether he wants to add to the house or buy something, advise him where to find things at a low price. This way she will see that you care about her and support her.</p> <p>When you share dreams and aspirations, make sure that your words do not give him the suspicion that you are proud, otherwise he will look for a way to distract you. Some of the things should be your secrets. When he asks you questions about something, give a longer answer that answers his questions. If there is something that does not satisfy you, speak clearly and humbly, solve the problem in a way that will not cause conflict. If your neighbor thinks that you are a high-class person, don't do things that confirm that idea. Stay humble at all times, don't pretend to be better than him. If this neighbor lies about you, ignore it and wait for someone else to point it out and take the step to talk about it.</p>

Table 3: Sesotho FAL 2012 summary writing text extracted from the February-March exam together with the corresponding Google translation of the Sesotho FAL 2012 summary writing exam text.

this version of the text, we have removed the sentence markers <utt> that were inserted during Sibeko and van Zaanen's (2022c) tokenization and sentence segmentation process. The text contains examples of figurative language. The Google Translate machine translation in table 3 indicates that at least all the words are successfully translated. Even so, instances of figurative language used in the Sesotho source text were translated out of context and new meanings were created.

The machine translations were post-edited to enable checking whether meaning influenced the readability of texts. Our translation corpus contains the original Sesotho texts, the original machine translations, and the human post-edited versions. The post-editing brief indicated that texts should not be changed unless meaning had been lost. As a result, in the human post-edited versions, machine translations such as liking the news were adapted to meaning-appropriate constructions such as nosy neighbours.

6 Conclusion

This article reported a work-in-progress PhD project. A survey of methods used for measuring text readability in low-resource languages indicated a prevalence of adapting classical readability metrics from high-resourced languages such as English. One of the common methods for adapting classical readability metrics was the use of translated texts between higher-resourced languages as helper languages and lower-resourced languages. Classical readability metrics use shallow textual features such as (i) the number of words, and (ii) the lengths of sentences, both of which can easily be counted, (iii) syllabic information for which we had to develop systems, and (iv) a frequency wordlist. Four resources were created and made available on the SADiLaR repository. One more resource is still under development. The resources include both gold-standard corpora and basic digital language resources such as syllabification systems. Finally, we identified the metrics we hope to adapt to Sesotho while also indicating the textual properties considered by each metric. At this point, we have put together most of the necessary tools for the identification and assessment of surface-level textual properties used in the nine readability metrics chosen. The main aim of the bigger study is to develop a platform for automated measurement of the readability of Sesotho texts. To this end, future works include the development of a list of frequently

used words in Sesotho which will then enable the adaptation of the Dale-Chall index to Sesotho. Furthermore, all nine metrics will be adapted and a web-based platform will be developed and made publicly accessible.

Acknowledgements

The doctoral project described in this article is completed at the North-West University in South Africa. It is supervised by Professor Menno van Zaanen of the South African Centre for Digital Language Resources.

References

- Alkaldi, W. 2022. Enhancing text readability using deep learning techniques. Université d'Ottawa/University of Ottawa.
- Andova, A. 2017. Assessment of text readability using statistical and machine learning approaches. University of Ljubljana.
- Anderson, J. 1983. Lix and Rix: Variations on a little-known readability index. *Journal of Reading*, 26(6):490–496.
- Barry, J. G. 1980. Computerized readability levels. *IEEE Transactions on Professional Communication*, 23(2): 88–90.
- Barry, J. G. and Stevenson, T. E. 1975. Using a computer to calculate the Dale-Chall formula. *Journal of Reading*, 19(3):218–222.
- Bendová, K. 2021. Using a parallel corpus to adapt the Flesch Reading Ease formula to Czech. *Journal of linguistics*, 72(2):477–487.
- Bendová, K. and Cinková, S. 2021. Adaptation of classic readability metrics to Czech. *Proceedings of the International Conference on Text, Speech, and Dialogue: 24th International Conference*, 159–171.
- Björnsson, C.-H. 1983. Readability of newspapers in 11 languages. *Reading Research Quarterly*, 480–497.
- Boles, C. D., Liu, Y., and November-Rider, D. 2016. Readability levels of dental patient education brochures. *American Dental Hygienists' Association*, 90(1):28–34.
- Chitja, M. 2010. *Phatlamantsoe ya Sesotho ya machaba*. Mazenod Publishers, Maseru, Lesotho.
- Coleman, M. and Liau, T.L. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.
- Collins-Thompson, K. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.
- Courtis, J. K. 1987. Fry, SMOG, Lix and Rix: Insinuations about corporate business communications. *The Journal of Business Communication*, 24(2):19–27.
- Dale, E. and Chall, J. S. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, 37–54.
- De Clercq, O. and Hoste, V. 2016. All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics*, 42(3):457–490.
- Demuth, C. 2007. Sesotho speech acquisition. In McLeod, S. *The international guide to speech acquisition*. 526–538. Thomson Delamar Learning: New York, USA.
- Dios, I. G. 2016. Readability assessment and automatic text simplification. The analysis of Basque complex Structures. University of the Basque Country.
- DuBay, W. H. 2004. The principles of readability. *Impact Information*, Costa Mesa: Online Submission, 1–76.
- Eleyan, D., Othman, A., and Eleyan, A. 2020. Enhancing software comments readability using Flesch Reading Ease score. *Information*, 11(9):430–455.

- Feng, L. 2010. Automatic readability assessment. New York: City University of New York.
- Filho, J. A. W., Wilkens, R. S., Zilio, L., Idiart, M., and Villavicencio, A. 2016. Crawling by readability level. In *Proceedings of the 12th International Conference on Computational Processing of the Portuguese Language*, Vol 1: 306–318.
- Flesch, R. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221–233.
- François, T. and Fairon, C. 2012. An “AI readability” formula for French as a foreign language. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, 466-477.
- Greenfield, J. 2004. Readability formulas for EFL. *Japan Association for Language Teaching*, 26(1):5–24.
- Guma, S. M. 1982. *An outline structure of Southern Sotho*. 2nd ed. Shooter and Shuter Publishers: Pietermaritzburg, South Africa.
- Gunning, R. 1952. *The technique of clear writing*. McGraw-Hill: New York.
- Gunning, R. 1969. The fog index after twenty years. *Journal of Business Communication*, 6(2):3-13.
- Gunning, T. 2003. The role of readability in today’s classrooms. *Topics in Language Disorders*, 23(3):175–189.
- Hartley, J. 2016. Is time up for the Flesch measure of reading ease? *Scientometrics*, 107(3):1523–1526.
- Heydari, P. 2012. The validity of some popular readability formulas. *Mediterranean Journal of Social Sciences*, 3(2):432–423.
- Janan, D. 2011. Towards a new model of readability. University of Warwick.
- Kasabwala, K., Agarwal, N., Hansberry, D. R., Baredes, S., and Eloy, J. A. 2012. Readability assessment of patient education materials from the American Academy of Otolaryngology—Head and Neck Surgery Foundation. *Otolaryngology—Head and Neck Surgery*, 147(3):466–471.
- Kaur, S., Kaur, K., and Kaur, P. 2018. The influence of text statistics and readability indices on measuring University websites. *International Journal of Advanced Research in Computer Science*, 9(1):403–414.
- Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. 1975. Derivation of new readability formulas (Automated Readability index, Fog count and Flesch Reading Ease formula) for navy enlisted personnel. Defense Technical Information Center: *Report*.
- Krige, D. and Reid, M. 2017. A pilot investigation into the readability of Sesotho health information pamphlets. *Communitas*, 22:113–123.
- Kondru, J. 2006. Using part of speech structure of text in the prediction of its readability. The University of Texas at Arlington.
- Lemeko, P. A. 2018. Diachronic investigation into current issues in language variation. A case of Sesotho language. Bloemfontein: Central University of Technology, Free State.
- Lesotho. 1993. The Constitution of Lesotho. Government Printer: Lesotho.
- Marupi, O. and Charamba, E. 2022. Revisiting the effects of –isms in the promotion, development, and revitalisation of indigenous languages in Zimbabwe: The position of Sesotho in Gwanda South, Zimbabwe. *Handbook of Research on Teaching in Multicultural and Multilingual Contexts*, 32–46.
- Mc Laughlin, G. H. 1969. SMOG grading – a new readability formula. *Journal of reading*, 12(8):639–646.
- Mohammed, L. A., and Aljaberi, M. A., Anmary, A. S. and Abdulkhaleq, M. 2022. Analysing English for Science and Technology reading texts using Flesch Reading Ease online formula: The preparation for academic reading. *International Conference on Emerging Technologies and Intelligent Systems*, 546–561.
- Mohasi, L., Mixdorff, E., and Niesler, T. 2011. An acoustic analysis of tone in Sesotho. *ICPhS*, 17–21.
- Newbold, N. 2013. *New Approaches for Text Readability*. United Kingdom: University of Surrey.
- Oborneva, I. V. 2006. Avtomatizirovannaya otsenka slozhnosti uchebnykh tekstov na osnove statisticheskikh parametrov [Automated estimation of the complexity of educational texts on the basis of statistical parameters]. RAS Institut sodержaniya i metodov obucheniya [RAS Institute of Content and Teaching Methods].

- Reid, M., Neil, M., Janse Van Rensburg-Bonthuyzen, E. 2019. Development of a Sesotho health literacy test in a South African context. *African Journal of Primary Health Care and Family Medicine*, 11(1):1–13.
- Roux, J. C. and Bosch, S. E. 2019. Preserving and developing indigenous languages in the South African context. *Proceedings of the Language Technologies for All*. European Language Resources Association. 97–100.
- Roux, J. C. and Ndinga-Koumba-Binza, S. African languages and human language technologies. 2019. *The Cambridge Handbook of African Linguistics*, 623–644.
- Sekere, N. B. 2004. Sociolinguistic variation in spoken and written Sesotho: A case study of speech varieties in Qwaqwa. University of South Africa, Pretoria.
- Sjöholm, J. 2012. Probability as readability: A new machine learning approach to readability assessment for written Swedish.
- Smith, E. A. and Senter, R. J. *Automated Readability index*. University of Cincinnati, Ohio.
- Sibeko, J. 2021. A comparative analysis of the linguistic complexity of Grade 12 English Home Language and English First Additional Language examination papers. *Per Linguam*, 37(2):50–64.
- Sibeko, J. and Setaka, M. 2022. An overview of Sesotho BLARK Content. *Journal of Digital Humanities Association of South Africa*, 4(2):1–11.
- Sibeko, J. and Van Zaanen, M. 2021. An analysis of readability metrics on English exam texts. *Journal of the Digital Humanities Association of Southern Africa*, 3(1):1–11.
- Sibeko, J. and Van Zaanen, M. 2022a. Sesotho syllabification systems. *Southern African Centre for Digital Language Resources*. Available at: <https://repo.sadilar.org/handle/20.500.12185/555> [Accessed: 3 Jan 2023].
- Sibeko, J. and Van Zaanen, M. 2022b. Raw and syllabified wordlist for Sesotho. *Southern African Centre for Digital Language Resources*. Available at: <https://repo.sadilar.org/handle/20.500.12185/556> [Accessed: 3 Jan 2023].
- Sibeko, J. and Van Zaanen, M. 2022c. Final year high school examination texts of South African Home and First Additional language subjects. *Southern African Centre for Digital Language Resources*. Available at: <https://repo.sadilar.org/handle/20.500.12185/568> [Accessed: 29 Dec. 2022].
- Solovyev, V., Ivanov, V., and Solnyshkina, M. I. 2018. Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics. *Journal of intelligent and fuzzy systems*, 5(34):3049–3058.
- Stocker, L. P. 1971. Increasing the precision of the Dale-Chall readability formula. *Reading Improvement*, 8(3):87.
- Tanprasert, T. and Kauchak, D. Flesch-Kincaid is not a text simplification evaluation metric. *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, 1–14.
- Thomas, G., Hartley, R. D., and Kincaid, J. P. 1975. Test-retest and inter-analyst reliability of the automated readability index, Flesch Reading Ease score, and the fog count. *Journal of Reading Behavior*, 7(2):149–154.
- Toyama, Y., Hiebert, E. H., and Pearson, P. D. 2017. An analysis of the text complexity of leveled passages in four popular classroom reading assessments. *Educational Assessment*, 22(3):139–170.
- Van Oosten, P., Tanghe, D., and Hoste, V. 2010. Towards an improved methodology for automated readability prediction. *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010)*, European Language Resources Association (ELRA). 775–782.
- Wilken, I., Gumede, T., Moors, C., and Calteaux, K. 2018. Human language technology audit 2018: Design considerations and methodology. *International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, 1–7.
- Wong, K. and Levi, J. R. 2017. Readability of pediatric otolaryngology information by children’s hospitals and academic institutions. *The Laryngoscope*, 127(4):E138–E144.
- Yan, X., Song, D., and Li, X. 2006. Concept-based document readability in domain specific information retrieval. *Proceedings of the 15th ACM international conference on Information and knowledge management*, 540–549.
- Zamanian, M. and Heydari, P. 2012. Readability of texts: State of the art. *Theory and Practice in Language Studies*, 2(1):43–53.

- Zhang, Y., Lin, N., and Jiang, S. 2019. A Study on syntactic complexity and text readability of ASEAN English news. *2019 International Conference on Asian Language Processing (IALP)*, 313–318.
- Zhou, S., Jeong, H., and Green, P.A 2017. How consistent are the best-known readability equations in estimating the readability of design standards? *IEEE Transactions on Professional Communication*, 6(1):97–111.

WebLicht-Batch – A Web-Based Interface for Batch Processing Large Input with the WebLicht Workflow Engine

Claus Zinn

Department of Linguistics
University of Tuebingen, Germany
claus.zinn@uni-tuebingen.de

Ben Campbell

Department of Linguistics
University of Tuebingen, Germany
ben.campbell@uni-tuebingen.de

Abstract

WebLicht is a workflow engine that gives researchers access to a well-inhabited space of natural language processing tools that can be combined into tool chains to perform complex natural language analyses. In this paper, we present WebLicht-Batch, a web-based interface to WebLicht’s chainer back-end. WebLicht-Batch helps users to automatically feed large input data, or input data of multiple files into WebLicht. It disassembles large input into smaller, more digestible sizes, feeds the resulting parts into WebLicht’s pipelining and execution engine, and then assembles the results of such processing into files that preserve the usual input-output dichotomy.

1 Introduction

WebLicht is a web-based application that allows users to easily create and execute tool chains for linguistic analysis. No software must be downloaded or installed as all computation is delegated to tools that WebLicht knows about and interacts with on users’ behalf (Hinrichs et al., 2010).

For a couple of reasons, WebLicht has a size limit on the data that users can upload for processing. First and foremost, WebLicht must take into account the analysis capabilities of the services it gives access to. While some services can cope with a large amount of data, others struggle with much less data to process. Second, WebLicht needs to keep the computation time of the services connected to WebLicht within a reasonable limit, and network-related socket timeouts need to be avoided, if possible. And third, but last, the output of the analyses can get rather large, but this is usually connected to the first two items.

In this paper, we present WebLicht-Batch, a browser-based service built upon the WebLicht backend that helps users to invoke WebLicht with large input. Our work also supports users that need to process a set of text files at once. Rather than submitting them manually to WebLicht, users can upload them as a collection archive so that WebLicht-Batch can process the collection item by item. Both usage scenarios are intertwined with each other in cases where a collection of files contains one or more large files.

2 Background

WebLicht is an execution environment for natural language processing pipelines. It uses a service-oriented architecture (SOA), where web services can be combined into processing chains. Chains are executed via sequential HTTP POST requests to services on the chain; here, the output of service n is the input to service $n + 1$ in the chain. Most services in WebLicht use Text Corpus Format (TCF)¹ as their input and output, and each service usually adds one or more annotation layer(s) to the result file.

Fig. 1 shows the main architecture of WebLicht. WebLicht makes use of a harvester to gather CMDI-based metadata of WebLicht-compatible web services from participating metadata repositories.² For the following discussion, take the Charniak parser (Charniak, 2000), which is addressable via a persistent identifier³ that points to the CMDI-based metadata description of the tool. Each service description obtained from such harvesting describes a service in terms of its name (e.g. “Charniak Parser +POS”), the

¹This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

²https://weblight.sfs.uni-tuebingen.de/weblightwiki/index.php/The_TCF_Format

³<https://weblight.sfs.uni-tuebingen.de/apps/harvester/resources/services>

⁴<http://hdl.handle.net/11022/0000-0000-8496-1>

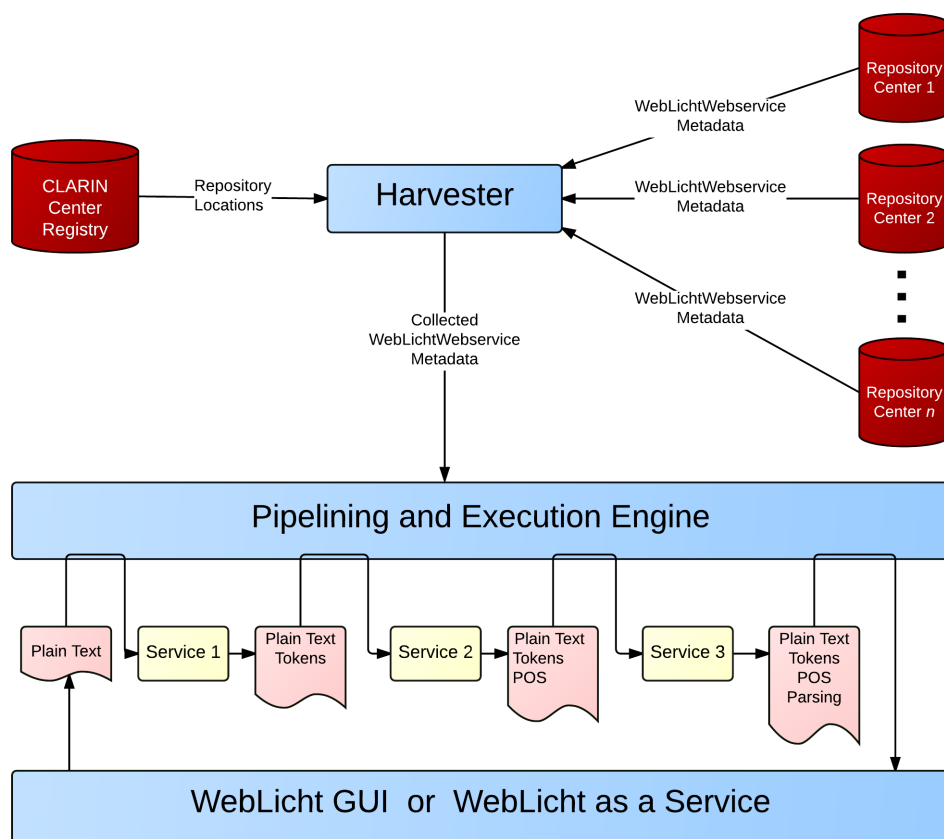


Figure 1: WebLicht's Architecture.

processing it performs (e.g., “BLLIP Parser is a statistical natural language parser including a generative constituent parser (first-stage) and discriminative maximum entropy reranker (second-stage). This service comes with the default model provided by BLLIP parser”), contact information, its life cycle status (e.g., “production”), and a WADL URL, which gives a service description in terms of the Web Application Description Language.⁴

Note that the WADL description only lists the service's endpoint⁵ Information that informs WebLicht's Pipelining and Execution Engine is encoded outside of WADL in the CMDI-based web service description. Fig. 2 visualizes this metadata using the CMD Orchestration Metadata Editing Tool (COMET).⁶ Here, it is specified that the input must be in English and complemented with TCF-based annotations for tokens and sentences, and that the output will add annotations layers for part-of-speech tags and parsing tagsets.⁷ Note, however, that none of the metadata fields specify the size of the input as tool selection or invocation constraint.

At the time of writing, WebLicht harvests all repositories known to the CLARIN Center Registry of which 27 repositories have WebLicht web service descriptions of 572 services.⁸ Frequently used services that are part of commonly-used NLP pipelines are installed and hosted directly on our institution-based servers, but most services run on many different servers in Germany and worldwide.

⁴<https://www.w3.org/Submission/wadl>

⁵For the Charniak parser this is the URL <http://weblight.sfs.uni-tuebingen.de/rws/parsers/service-charniak/annotate/parse>, together with the mediaType of the request as well as its response, usually of type “text/tcf+xml”.

⁶<https://weblight.sfs.uni-tuebingen.de/comet>

⁷See the appendix for an example of a TCF-based input representation.

⁸The harvester has an update interval of 2 hours, and hence the overall tool range of WebLicht may change that frequently, see the harvester report at <https://weblight.sfs.uni-tuebingen.de/harvester/resources/report>.

Orchestration Information			
Input		Output	
Name	URL Argument	Name	Input Reference
lang		parsing.tagset	
en		penntb	
sentences		postags.tagset	
tokens		penntb	
type		+ Add Feature	
text/tcf+xml		<input type="checkbox"/> Replaces Input	
version			
0.4			
5			

Figure 2: Orchestration Metadata from Charniak’s Parser.

The WebLicht GUI⁹ provides users with a web-based interface to upload their data and get it processed by their NLP pipeline of choice. In WebLicht’s Easy Mode, users can choose among pre-defined processing chains that match often-used linguistic pipelines.¹⁰ In Advanced Mode, users are supported to build *permissible* tool chains to customise or finetune the processing for the intricacies of the task at hand.

WebLicht as a Service (WaaS) is a REST service that executes WebLicht chains.¹¹ Unlike the WebLicht web application, WaaS does not require a browser, and hence prevents browser-specific issues from arising such as file size upload limits. Also, it does not impose on users to perform the rather mundane task of actioning a GUI to get processing started. With WaaS, users can run chains from their UNIX shell, scripts, or programs. Once users have defined a chain in the WebLicht browser interface, they can download the chain, and then they can execute a HTTP POST request with the multipart/form-data encoding to invoke WaaS with the chain in question and the input data.¹²

Note, however, that WaaS is not always the solution to process a single large file, or a collection of smaller files. First, there are some services in the WebLicht tool space that cannot handle large files *per se*. Once they fail on large input, the entire processing chain fails and no output is returned to users. In this case, users will need to manually split the input into smaller entities, get them processed one by one, and assemble the individual results into a compound entity. Also, some users are not comfortable mechanising such enterprise with a program script.

3 WebLicht-Batch

Fig. 3 depicts the central idea of WebLicht-Batch. A large plain text input file is split into multiple smaller files at sentence boundaries. Each individual file is then sent to WebLicht’s pipelining and execution engine that processes the file with the NLP pipeline chosen by the user. The result of processing each file is captured in TCF format; they are then assembled to form a compound TCF-based result file. When users submit a ZIP file to WebLicht-Batch, each file in the archive is processed in the same manner. In addition, a ZIP file is constructed that contains the results of processing the individual files.

WebLicht-Batch makes use of WebLicht’s pipelining and execution engine and provides, in addition,

⁹<https://weblight.sfs.uni-tuebingen.de/weblight>

¹⁰See the appendix for an easy-chain that makes use of the Charniak parser.

¹¹<https://weblight.sfs.uni-tuebingen.de/WaaS>

¹²For instance, by executing curl commands such as `curl -X POST -F chains=@chains.xml -F content=@inputFile -F apikey=apiKey https://weblight.sfs.uni-tuebingen.de/WaaS/api/1.0/chain/process > result` – Note that an apiKey must be acquired from the WaaS website.

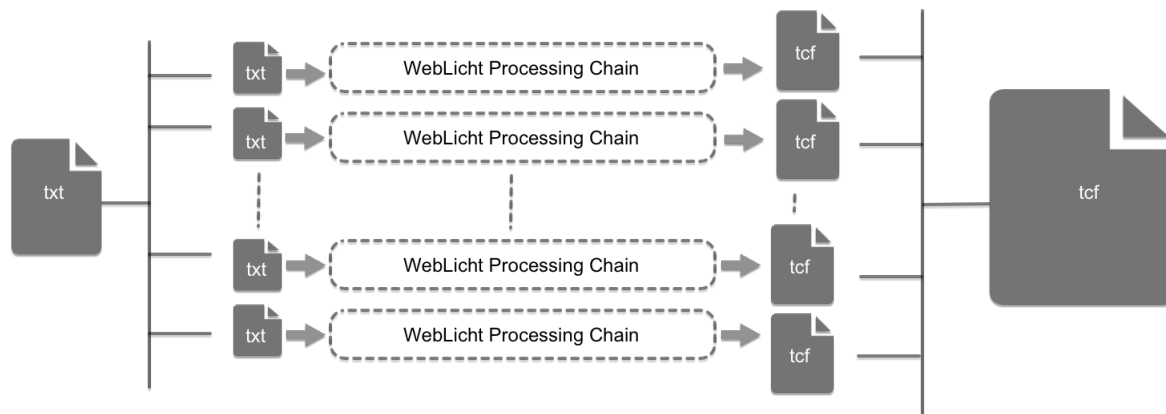


Figure 3: WebLicht-Batch – Central Idea.

an API to upload a file (in plain text format, or ZIP format), to upload a chain, to start (or cancel) the batch process, to get processing information, and to retrieve the result file. The front-end of WebLicht-Batch makes use of this API and guides users through the overall process. WebLicht-Batch, hence, joins the WebLicht GUI and WebLicht As a Service as a third “user” of the Pipeling and Execution Engine.

WebLicht-Batch Front-End. Fig. 4 depicts the main GUI of WebLicht-Batch. Here, users can upload a single file, which can either be a plain text file, or a collection thereof being archived in ZIP format. Users then select the language of the text file(s) they want to process and also the processing chain they would like to run on the file(s). WebLicht-Batch gives access to all easy-chains offered by the WebLicht GUI, but users can also upload their own processing chain.¹³ When users then press the “Start Processing” button, the batch-processing is started. Also, a user-specific key (“userkey”) is generated that users are encouraged to copy to their clipboard. The user-key allows users to inspect the task status at a later time, even if they closed the browser tab in the mean time.

The figure also depicts the task progress for a plain text file that we have given to WebLicht-Batch. The text file has a size of approximately 200 kilobytes and was split into a batch of three files. For each of the three files, a table lists the progress, including the service that is currently run for each batch item. In our example, the last file has completed processing in WebLicht’s pipelining and execution engine whereas item 1 and 2 are still being processed by Charniak’s parser.

WebLicht-Batch Back-End. The basic requirements for any WebLicht-Batch task are that there is a valid text or ZIP file, and a valid WebLicht chain file. After verifying that the chain file is valid, it is next determined whether the input file is a text or a ZIP. If it is neither, an error is returned.¹⁴

The first step is the splitting of the original input text file into 100KB chunks, a size that most WebLicht services are comfortable with. This is somewhat of a “chicken and egg” problem since, in order to split the file, it is necessary to use NLP tools which can perform this splitting, but which we do not want to feed too large of a file into, which requires the files to be split before sending them into the file splitter. In order to resolve this issue, we make use of the UDPipe tokeniser and sentence splitter (Straka and Straková, 2017) and feed in 100KB sized chunks – this size was chosen for the sake of convenience, as it is the same as the chunk size we use to perform batch processing.¹⁵ Splitting the file at 100KB results in

¹³Processing chains are represented in an XML-based format. Users are advised to define, test, and download them using WebLicht’s Advanced Mode. For a chain example, see the appendix.

¹⁴The design rationale of WebLicht-Batch allows users to process files of arbitrary size. While there hence no technical reason to limit file uploads, there is a practical one, fairness. Other users with smaller inputs should get access to WebLicht’s space of NLP services and not be blocked by power users wanting to process overly large inputs. To take this fairness constraint into account, we restricted the maximum allowable size for any single text file to 2.5MB, while the maximum size for a ZIP file is 50 MB.

¹⁵Defining the threshold of 100 kilobytes is informed by our long-time experience working with WebLicht. The performance of some services in the WebLicht space of services degrade significantly (or even fail) when given inputs larger than 100 kilobytes.

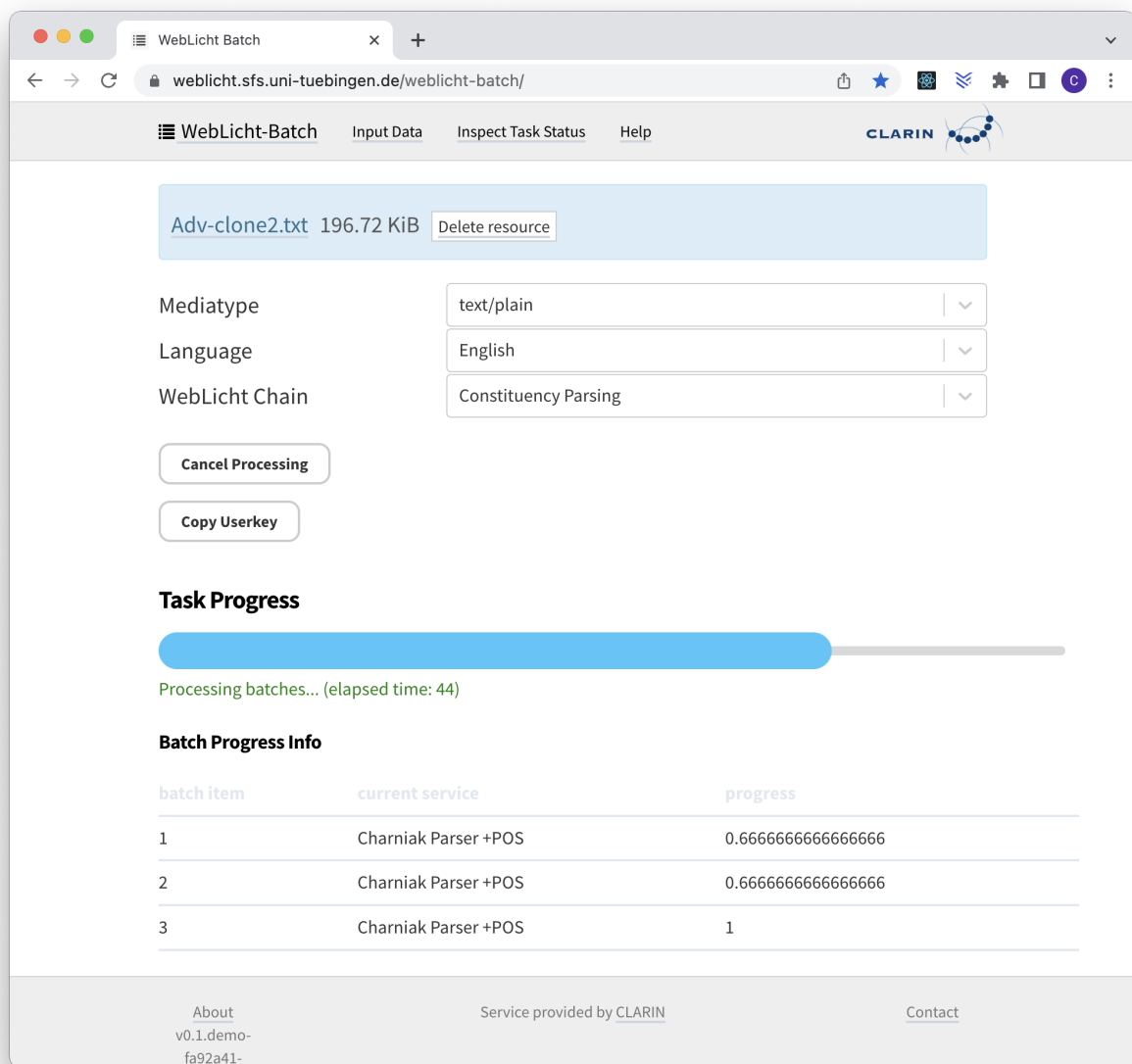


Figure 4: The WebLicht-Batch GUI.

text files which are split at arbitrary points in the final sentence. We start by sending the first chunk into the UDPipe tokeniser and sentence splitter, and assume that the last sentence of the output is incomplete, and then remove this sentence from the output of the first chunk, then add it to the beginning of the next chunk, which is then fed into UDPipe. This process is repeated until all chunks have been split into sentences. These chunks are then stored on the server to await further processing.

Next, we use the WebLicht chainer to process each chunk. At the time of this writing, the batch processor allows four chunks to be processed simultaneously as batches, which should allow a reasonable tradeoff between parallelism, and thus overall processing speed, and not overloading any of the services. Progress data, including which service of the chain is currently processing the chunk is constantly collected and sent to the frontend. If there is a failure in the processing of a chunk at any point, it is attempted to again run the chain on the chunk which failed. After three failures in a row, it is considered a failed batch and the entire task is considered to have failed.

If all batches succeed, the resulting TCF output files are then combined into one large TCF file. This is a complex process which involves manipulating the annotation layers for each TCF output file in order

to ensure that the token ids for each token are correct for each annotation layer. If this combining is successful, a download link for the resulting file is sent to the frontend.

For a ZIP archive of plain text files, each file is processed as described above. The resulting TCF files are then packed into a ZIP file of which the download link is sent to the frontend. If the processing of any file in the archive fails, the entire processing is not considered to have failed. Rather, a list of files which have failed is kept and processing of the other files in the archive continues. After processing of all files is complete – whether some have failed or not – there are a number of lists which are stored on the server as files. These include a “failed” list, a list of files whose processing failed at some point, a “tooLarge” list, a list of files which could not be processed due to being larger than the 2.5MB limit, and an “invalidFormat” list, a list of files which could not be processed due to not being plain text. These list files are also packed into the output ZIP file which the user can download.

WebLicht-Batch has been integrated with CMDI Explorer (Arnold et al., 2020), a web-based tool that helps users explore collections that are described with CMDI. In CMDI Explorer, users can select plain text files in the collection tree, request the generation of a ZIP file to bundle them, and send the archive to WebLicht-Batch for further processing. WebLicht-Batch has also been integrated with the Language Resource Switchboard (Zinn, 2018). When users upload a ZIP file to the Switchboard, WebLicht-Batch is shown as applicable tool. Once started, users are left to specify the common language of the text files and a WebLicht processing chain.

4 Discussion and Future Work

WebLicht As a Service is a REST-based API where access to WebLicht’s pipeling and execution service is given via HTTP requests, and hence, callable from Java, Python, and other programming languages. By its design, it addresses the issue of browser-depended timeouts. Script-based by nature, it allows developers to invoke the script whenever they need it, or when they think WebLicht’s army of services is idle rather than busy. Also, it is straightforward to invoke the script on a set of files, which is rather clumsy to achieve in the WebLicht GUI. Note, however, that for large input, the WebLicht As a Service approach delegates the responsibility of file splitting at sentence boundaries and the combination of individual TCF files into a compound TCF to its users. Both file splitting and results re-combination are non-trivial tasks that many users may not want to perform themselves. Those users will welcome WebLicht-Batch.

Apart from WaaS, we know of only one other application that addresses the processing of large data with WebLicht. But rather than splitting large input into more digestible chunks, it aimed at placing WebLicht services and the data they need to process into a shielded, high-performance environment – for big data (and also for sensitive data), it is better to move the tools to the data rather than having the data travel to the tools. In (Zinn et al., 2018), the Generic Execution Framework (GEF, stemming from the EUDAT project) has been used to provide such environments. WebLicht services were installed in a so-called GEF environment with direct access to the data to be processed. A development version of WebLicht was built that had access to the GEF environment; and when users uploaded data to this version of WebLicht, the data was transferred to the location that also hosted the services.

The installation of GEF-ified services gives GEF maintainers the opportunity to preselect NLP services that can either cope with large data, or install many instances of the same service to handle many processing requests in parallel. While the installation of such purpose-built computing environments for the processing tasks at hand is costly, it helps minimising users’ waiting times or processing errors. GEF itself was built using Docker software containerisation technology, was seen as part the EUDAT Collaborative Data Infrastructure, but has never entered production mode; for more details, see <https://github.com/EUDAT-GEF/GEF>.

There are a number of issues that we would like to tackle in the future. Most services that are part of WebLicht’s easy-chains are installed locally at institutional servers using Docker technology. For large input, we would like to investigate how to use Docker to spawn new workers of a given service on the fly giving a rising demand from WebLicht-Batch users. However, care must be taken to not overload individual services. A large WebLicht-Batch process could block regular WebLicht GUI users from getting their (smallish) input processed in time. Here, batch processing may want to postpone heavy processing

to a point in time where Docker-based services are idle. Here, we may want to give users a scheduling option, where users are told estimated processing times depending on the time slots they choose.

From a practical perspective it is usually one service per chain that causes a bottleneck; this is usually a service offering constituency or dependency parsing, a rather complex process compared to tokenisation or part-of-speech-tagging. Here, we need to investigate whether complex processes should be given more CPU power and memory, or more workers by default, than simpler analyses.

To gain a better understanding of service use and performance, it would be useful to gather certain statistics. For example, which services are most used, which take the longest to process (and thus are more likely to cause bottlenecks), and which can process the most chunks of data in parallel. All this data could be used to customize the processing of each task in order to maximize speed and efficiency, rather than the current “one size fits all” approach to handling tasks.

In addition, more work is required to better understand the trade-off between the item sizes within a batch and the cost of splitting input into smaller chunks and the reassembling of individual results into a compound result. Also, the processing chain selected by WebLicht-Batch users should be taken into account. Chains without bottleneck services might profit from larger rather than smaller chunk splitting.

Most WebLicht services (usually not being part of easy-chains) are installed outside the control of WebLicht developers. Given the overall architecture of WebLicht and its few hundreds of services that are distributed over many different servers, batch design and task scheduling is all but trivial.

Another improvement that could make WebLicht-Batch more useful to a wide variety of users would be increasing the maximum size of individual files allowed for upload. As of this writing the maximum size of a text file that can be uploaded is 2.5 MB. Apart from fairness considerations (see footnote 14), this is done due to the fact that the output files are often orders of magnitude larger than the input files, and we have a limitation on the size of the output files that can be downloaded of about 2 GB. This could be accomplished during the output file combination stage by combining the output TCF files into combined output files of less than 2 GB, and allowing the user to download multiple output files.

As of this writing it is only possible to upload text files for processing. But WebLicht-Batch could be further improved by allowing the upload of TCF files. This would present a technical challenge as it would be more difficult to split a TCF file than it is to split a text file. The reason for this is that the individual layers would have to be split and care would have to be taken to ensure that each layer is split at the same point in the text in order for the individual chunks to be processed properly. However, as WebLicht allows the upload of TCF and not just text files, it would be a good idea to add this at some point if it is technically feasible.

Finally, it could also be useful to include a link to our TüNDRA tool, which allows the visualization of TCF and CoNLL-U files.¹⁶ The user would be able to click the link and have their output visualised there. This can already be done with WebLicht, so it would make sense if this option were available for WebLicht-Batch too. One issue is that TüNDRA has a file upload size limit of 50 MB, so it may be a good idea to include an option prior to processing of having output file sizes of at most 50 MB, rather than having everything bundled into one TCF file (or multiple files of up to 2 GB, as discussed above).

5 Conclusion

In this paper, we have presented WebLicht-Batch, a browser-based application that supports users in feeding large files, or a ZIP archive of files into WebLicht. We believe that WebLicht-Batch is a good addition to the WebLicht family of tools. It complements our WebLicht GUI and WebLicht as a Service (Waas) software, relieving users from the burden of submitting many files of a collection one by one, or by splitting large input that WebLicht services fail to process into smaller, more manageable chunks. There is ample potential to improve the quality of batch processing the input, but it is a non-trivial task as it must be informed by gathering performance statistics from a highly distributed tool landscape.

We invite all readers to test, play around, and use the service, which is available at <https://weblicht.sfs.uni-tuebingen.de/weblicht-batch>. Feedback is highly welcome.

¹⁶<https://weblicht.sfs.uni-tuebingen.de/Tundra>

6 Acknowledgements

The work on WebLicht-Batch has been funded by the SSHOC project (Social Sciences & Humanities Open Cloud), a Horizon 2020 EU framework programme, project number 823782. Development of WebLicht started in October 2008 as part of the BMBF-funded D-SPIN project, the predecessor project of CLARIN-D, and continued through the CLARIN umbrella.

References

- Denis Arnold, Ben Campbell, Thomas Eckart, Bernhard Fisseni, Thorsten Trippel, and Claus Zinn. 2020. CMDI Explorer. In Costanza Navarretta and Maria Eskevich, editors, *Selected Papers from the CLARIN Annual Conference 2020*, volume 180 of *Linköping Electronic Conference Proceedings*, pages 8–15. Linköping University Electronic Press.
- Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *6th Applied Natural Language Processing Conference, ANLP 2000, Seattle, Washington, USA, April 29 - May 4, 2000*, pages 132–139. ACL.
- Marie Hinrichs, Thomas Zastrow, and Erhard W. Hinrichs. 2010. Weblicht: Web-based LRT services in a distributed escience infrastructure. In Nicoletta Calzolari et al., editor, *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta*. ELRA.
- Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Claus Zinn, Wei Qui, Marie Hinrichs, Emanuel Dima, and Alexandr Chernov. 2018. Handling Big Data and Sensitive Data Using EUDAT’s Generic Execution Framework and the Weblicht Workflow Engine. In Nicoletta Calzolari et al., editor, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan*. ELRA.
- Claus Zinn. 2018. The Language Resource Switchboard. *Computational Linguistics*, 44(4):631–639.

Appendix

Fig. 5 depicts a TCF fragment where the annotation layers for tokens, sentences, and part-of-speech tags have been added to the one-sentence input shown in the `tc:text` tag. Fig. 6 depicts a WebLicht easy-chain for constituency parsing for English. It contains a service that converts plain text input into TCF, the Stanford tokeniser, which performs tokenisation and sentence splitting, and the aforementioned Charniak parser. Each PID points to the CMDI-based service description of the tool.

```

2 <md:MetaData xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:cmd="http://www.clarin.eu/cmd/"
3 <TextCorpus xmlns="http://www.dspin.de/data/textcorpus" lang="en">
4 <tc:text xmlns:tc="http://www.dspin.de/data/textcorpus">YOU don't know about me without you have
5 read a book by the name of The Adventures of Tom Sawyer; but that ain't no matter.</tc:text>
6 <tc:tokens xmlns:tc="http://www.dspin.de/data/textcorpus">
7 <tc:token ID="t_0">YOU</tc:token>
8 <tc:token ID="t_1">do</tc:token>
9 ...
10 <tc:token ID="t_27">matter</tc:token>
11 <tc:token ID="t_28">.</tc:token>
12 </tc:tokens>
13 <tc:sentences xmlns:tc="http://www.dspin.de/data/textcorpus">
14 <tc:sentence tokenIDs="t_0 t_1 ... t_27 t_28"/>
15 </tc:sentences>
16 <tc:POStags xmlns:tc="http://www.dspin.de/data/textcorpus" tagset="penntb">
17 <tc:tag tokenIDs="t_0">PRP</tc:tag>
18 <tc:tag tokenIDs="t_1">VBP</tc:tag>
19 ...
20 <tc:tag tokenIDs="t_27">NN</tc:tag>
21 <tc:tag tokenIDs="t_28">.</tc:tag>
22 </tc:POStags>

```

Figure 5: An abridged TCF example for the Representation of input text.

```

1 <cmd:CLARIN-D xmlns:cmd="http://www.clarin.eu/cmd/1" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
2 xsi:schemaLocation=
3 "http://www.clarin.eu/cmd/1 http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1320657629644/xsd">
4 <cmd:chains>
5 <cmd:CMD CMDVersion="1.2">
6 <cmd:Resources> [4 lines]
7 <cmd:Components>
8 <cmd:WebServiceToolChain>
9 <cmd:GeneralInfo>
10 <cmd:Descriptions>
11 <cmd:Description/>
12 </cmd:Descriptions>
13 <cmd:ResourceName>myChain</cmd:ResourceName>
14 <cmd:ResourceClass>Toolchain</cmd:ResourceClass>
15 </cmd:GeneralInfo>
16 <cmd:Toolchain>
17 <cmd:ToolInChain>
18 <cmd:PID>http://hdl.handle.net/11858/00-1778-0000-0004-BA56-7</cmd:PID>
19 <cmd:Parameter value="en" name="lang"/>
20 <cmd:Parameter value="text/plain" name="type"/>
21 <cmd:Parameter value="0.4" name="version"/>
22 </cmd:ToolInChain>
23 <cmd:ToolInChain>
24 <cmd:PID>http://hdl.handle.net/11022/0000-0000-2518-C</cmd:PID>
25 <cmd:Parameter value="0.4" name="version"/>
26 </cmd:ToolInChain>
27 <cmd:ToolInChain>
28 <cmd:PID>http://hdl.handle.net/11022/0000-0000-8496-1</cmd:PID>
29 <cmd:Parameter value="0.4" name="version"/>
30 </cmd:ToolInChain>
31 </cmd:Toolchain>
32 </cmd:WebServiceToolChain>
33 </cmd:Components>
34 </cmd:CMD>
35 </cmd:chains>
36 </cmd:CLARIN-D>

```

Figure 6: A WebLicht easy-chain for constituency parsing for English text input.